

# Dynamic Repair and Maintenance of Heterogeneous Machines Dispersed on a Network: A Rollout Method for Online Reinforcement Learning

Dongnuan Tian

Department of Management Science, Lancaster University Management School, Lancaster University, Lancaster  
LA1 4YW, United Kingdom. Email: d.tian2@lancaster.ac.uk.

Rob Shone\*

Department of Management Science, Lancaster University Management School, Lancaster University, Lancaster  
LA1 4YW, United Kingdom. Email: r.shone@lancaster.ac.uk.

---

## Abstract

We consider a problem in which a single repairer is responsible for the maintenance and repair of a collection of machines, positioned at different locations on a network of nodes and edges. Machines deteriorate according to stochastic processes and incur increasing costs as they approach complete failure. The times needed for repairs to be performed, and the amounts of time needed for the repairer to switch between different machines, are random and machine-dependent. The problem is formulated as a Markov decision process (MDP) in which the objective is to minimize long-run average costs. We prove the equivalence of an alternative formulation based on rewards and use this to develop an index heuristic policy, which is shown to be optimal in certain special cases. We then use rollout-based reinforcement learning techniques to develop a novel online policy improvement (OPI) approach, which uses the index heuristic as a base policy and also as an insurance option at decision epochs where the best action cannot be selected with sufficient confidence. Results from extensive numerical experiments, involving randomly-generated network layouts and parameter values, show that the OPI heuristic is able to achieve close-to-optimal performance in fast-changing systems with state transitions occurring 100 times per second, suggesting that it is suitable for online implementation.

*Keywords:* Repair and maintenance, policy improvement, reinforcement learning, rollout

---

\*Corresponding author

## 1 Introduction

A classical problem in operations research concerns the dynamic scheduling of maintenance and repair operations for a collection of machines that are prone to failure. These problems fall within the broader framework of reliability modeling, which is concerned with the performance of systems that degrade or fail over time due to random influences. Related problems have been studied extensively due to their numerous applications in computer systems, telecommunications, transportation infrastructure, manufacturing systems, and other areas (see, e.g., Seyedshohadaie et al. (2010); Ahmad et al. (2017); Poonia (2021); Geurtsen et al. (2023)). In the context of machine failures, it is often assumed that each machine has a finite number of possible states, ranging from the pristine or ‘as-good-as-new’ state (in which it functions at its optimal performance level) to the ‘failed’ state (in which it performs at its worst level, or does not function at all). If a machine is left unattended, a *failure process* causes it to pass through a sequence of increasingly undesirable states until it eventually reaches the failed state. The times of state transitions are random, so that the speed of the failure process is unpredictable. The system controller has a set of resources available (in the form of repair engineers, for example) to stall or reverse the failure process by effecting transitions in the opposite direction, i.e. towards the pristine state, and must allocate these resources efficiently over time in order to optimize a pre-specified measure of performance. These types of model assumptions, which combine stochastic degradation with resource-constrained control, have been extensively studied in the literature (see, e.g., Nino-Mora (2001); Glazebrook et al. (2005); Abbou and Makis (2019)).

The dynamic maintenance and repair problem can be extended and generalized in various ways. For example, one can assume that different kinds of maintenance and repair operations (known as ‘rate-modifying activities’) are available (Lee and Lin (2001); Woo et al. (2017); Mosheiov and Oron (2021)), or allow machine failures to be governed by multiple dependent failure processes (Yousefi et al. (2020); Ogunfowora and Najjaran (2023); Ward et al. (2024)). In this paper we consider a version of the problem that, to the best of our knowledge, is original and enables insightful analyses based on a network model. We assume that there is a single repair operative on duty (referred to throughout the paper as a ‘repairer’) who bears sole responsibility for the maintenance and repair of  $m \geq 2$  machines. Although the repairer can switch from repairing one machine to another, this cannot be done instantaneously, and the decision-making process should therefore take into account the time required to re-allocate effort from one machine to another in addition to the time needed to perform repairs. Indeed, the machines might be in different physical locations, in which case some travel time would be needed to switch from one machine to another. Alternatively, even if travel times are negligible, there might be setup requirements for repairing individual machines, so that the repairer must expend some time to perform the necessary preliminary tasks before commencing the next repair. We refer to the time needed to re-allocate effort from one machine to another as a ‘switching time’, and introduce further complexity by allowing these switching times to be interruptible, so that the repairer may begin the process of switching to one machine but then take another action (such as changing course and switching to a different machine) before

completing the switch.

To motivate our network-based model, Figure 1 shows an example with 4 machines ( $m = 4$ ). The machines are represented as gray-colored nodes in the network, labeled 1, 2, 3 and 4. At any given time, the repairer occupies one node and can choose either to remain at the current node or attempt to switch to one of the adjacent nodes (connected by edges to the current node). In order to repair a particular machine, the repairer must occupy the corresponding node. Thus, in order to switch from repairing one machine to another, it must traverse an appropriate path through the network. The white-colored nodes represent intermediate ‘stages’ that must be completed during a switch. For example, to switch from machine 1 to 3, the repairer would need to pass through nodes 5, 6 and 7. However, the direction of movement can be changed at any time, so the repairer might follow a path such as (1, 5, 6, 5, 2). This would represent a switch from machine 1 to 3 that gets ‘interrupted’ in favor of a switch to machine 2 instead. By designing the network appropriately, we can model intricate relationships between the travel times or setup requirements of different machines, including the effects on switching times when the repairer makes partial progress in switching from one machine to another. In Section 3 we provide details of how the system can be modeled as a Markov decision process (MDP).

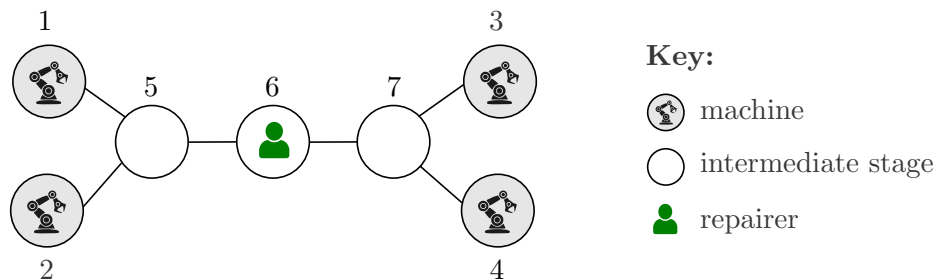


Figure 1: A network with 4 machines, 3 intermediate stages and a single repairer. Gray-colored nodes represent machines, and white-colored nodes represent intermediate stages.

One study of particular relevance to ours is that of Duenyas and Van Oyen (1996), which considers the dynamic assignment of a single server in a system of parallel, heterogeneous queues. Although their model is presented in a different context from ours (i.e. servicing of customers in a queueing system rather than repair of machines), there are similarities in its mathematical formulation, including the assumption of Poisson customer arrivals (analogous to machine degradations in our model) and the incorporation of random switching times. In our recent work (Tian and Shone (2026)), we have considered an extension of the job scheduling problem in Duenyas and Van Oyen (1996) that allows the server to interrupt processing and switching times and also uses a network formulation to encode the amounts of effort needed to switch between different tasks. The formulation in Tian and Shone (2026) has similarities to the one that we consider in the current paper, but there are two main differences. Firstly, in Tian and Shone (2026) it is assumed that there can be any number of jobs waiting to be processed at a demand point and this leads to

the formulation of an MDP with an infinite state space, in which system stability is an important consideration. In this paper we obtain a finite-state MDP because the set of possible conditions for any machine  $i$  is finite. While this removes the need to consider questions involving stability, it introduces other complications related to multichain policies and irregular value functions that we elaborate on in Sections 4 and 5. Secondly, in Tian and Shone (2026) it is assumed that the cost of having  $k$  jobs waiting at a particular demand point is linear in  $k$ , but in this paper we allow the cost function  $f(k)$  to be more general, as the effect on a machine's performance may be non-linear in the number of degradation events that has occurred. We also note that the solution methods in Tian and Shone (2026) are focused on index policies, whereas in this paper we begin by developing an index policy but subsequently use an altered version of this as a base policy in a novel online policy improvement algorithm.

The main contributions of our paper are as follows:

- We consider an original model of a repair and maintenance problem in which repair and switching times not only follow random distributions but are also interruptible. In addition, our proposed network formulation allows switching times to depend not only on the machines involved in the switch, but also on the amount of progress made during any switches that have been interrupted.
- We develop an index-based heuristic for choosing which machine to switch to next, with decisions being continuously re-calculated during repair and switching times in order to allow interruptions. In order to develop the heuristic, we first prove an equivalence between two cost formulations in the underlying MDP formulation. We then prove that the index heuristic is optimal in certain special cases of our problem.
- In order to obtain a stronger-performing heuristic policy we apply reinforcement learning (RL) techniques and approximate the value function for the index heuristic. We then use a rollout method, which we refer to as online policy improvement (OPI), to search for improved decisions during online implementation. Our OPI method includes a novel safety mechanism, based on reliability weights, which mandates that the decision given by the index policy is followed in situations where the best action cannot be identified with sufficient confidence.
- We present the results of computational experiments with randomly-generated network layouts and parameter values in order to test the performances of our heuristics and investigate the effects of adjusting key parameter values.

Section 2 provides a review of relevant literature, including a discussion of how our model could be applied in different application settings. Section 3 provides details of our mathematical formulation. Section 4 introduces our index heuristic and also includes proofs of its optimality in certain special cases. Section 5 explains the use of online policy improvement for obtaining a stronger-performing heuristic policy. Section 6 provides details of results from our numerical experiments.

Finally, our concluding remarks are given in Section 7.

## 2 Literature review

As discussed in Section 1, our study builds upon the previous literature on repair and maintenance problems. In particular, it introduces a network formulation to model travel (or setup) times between heterogeneous machines that degrade according to stochastic processes. However, our mathematical model and proposed solution approaches can also be applied more generally to other real-world settings in which limited resources are allocated dynamically to assets that degrade randomly over time. For example, we can consider network security problems in which connected devices are vulnerable to attacks or other disruptions, healthcare settings in which patient conditions worsen without effective intervention, or infrastructure monitoring systems in which bridges or roads are classified into discrete condition grades based on inspection criteria. In this section we summarize important contributions to the literature within each of these areas.

Within the maintenance optimization literature, foundational survey articles by Cho and Parlar (1991), Dekker et al. (1997) and Nicolai and Dekker (2008) have established the need for multi-component models with stochastic degradation processes and limited maintenance resources. In the classical literature, one can find several articles that (like our paper) formulate MDPs for maintenance decisions; see, for example, Berg (1976); Ohashi (1982); Özekici (1988); Dekker and Roelvink (1995); Wijnmalen and Hontelez (1997). These papers use dynamic programming methods to establish structural properties of optimal policies or develop heuristics in different settings, but they do not include the spatial features of the problem that we consider. More recently, several studies have modeled repair and maintenance problems using restless bandit formulations, based on the principle that only a certain subset of machines can be attended to at any given time, and those that are unattended are liable to change state (i.e. degrade) rather than remain in the same condition. Glazebrook et al. (2005) considered a discrete-time formulation with multiple repairers and used the Whittle's index method (see Whittle (1988)) to derive index heuristics in order to minimize expected discounted costs. Abbou and Makis (2019) also used Whittle indices to minimize discounted costs but proposed a group maintenance formulation, in which each 'bandit' is a production facility consisting of several machines with different possible states. Ruiz-Hernandez et al. (2020) considered a scenario with *imperfect* maintenance, in which repairs return a machine to an uncertain state of wear rather than the pristine state, and also used Whittle indices with discounted costs. Ayesta et al. (2021) developed computational methods based on general expressions for Whittle indices in Markovian bandit models and, as a special case, considered a machine repairman problem in continuous time under the average cost criterion. Demirci et al. (2025) have also applied the Whittle indices approach to a capacity-constrained maintenance problem. The restless bandits approach has some advantages in that, by considering a decomposition of the prob-

lem based on Lagrangian relaxation, one can easily extend the analyses to the case of multiple repairers. On the other hand, the aforementioned studies do not consider switching times (interruptible or otherwise) and indeed bandit-type models become much less tractable in general when the assumption of immediate, costless switching is removed.

In network security applications, our paper’s paradigm of stochastically degrading machines requiring resource allocation extends naturally to systems of connected devices, where vulnerabilities accumulate over time and must be addressed through patching, monitoring, or other defensive measures. Many examples of such problems can be found in the area of cybersecurity. Arora et al. (2010), Ioannidis et al. (2012) Dey et al. (2015) and Fielder et al. (2016) all develop stochastic models for information systems in which security risks intensify over time, although they do not use MDP formulations like ours. Żebrowski et al. (2022) consider the problem of mitigating cyber threats to cyber-physical systems and propose a Bayesian framework, which allows for different attack scenarios, to optimize allocation of resources across vulnerable components. Mookerjee and Samuel (2023) develop a stochastic dynamic programming formulation for a problem with partially observable vulnerabilities, in which decisions must be made as to when the system will be inspected. They show that optimal policies follow a threshold structure, with one (lower) threshold associated with the decision to inspect a vulnerability before fixing the problem and another (higher) threshold for implementing a fix directly. Medina et al. (2025) use stochastic optimization via attack graphs to construct integrated cybersecurity systems, employing two-stage stochastic programming for countermeasure allocation under budget constraints. Luo et al. (2025) propose an automatic detection model, formulated as an MDP, that uses automatic learning strategies to respond to evolving attack patterns and improve detection accuracy. Within the cybersecurity literature, many authors also study partially observable stochastic games (POSGs) for patch strategy optimization, based on the principle that defenders cannot fully observe attacker states or intentions, and may be unaware of the full environmental state. Recent work by Horák et al. (2023) has provided a comprehensive theoretical analysis of zero-sum one-sided POSGs, in which only one agent has imperfect information; however, extending these results to more general multi-agent settings remains a difficult task. Deep reinforcement learning (DRL) methods can be applied to security games with high-dimensional state spaces (Zhang et al. (2021)), but these approaches face computational challenges, including difficulties in converging to stable solutions (Hernandez-Leal et al. (2019)). Our paper assumes that machine degradation occurs through exogenous stochastic processes rather than strategic adversarial behavior, but we regard the consideration of strategic attack patterns as a promising direction for future research.

In the field of healthcare applications, the home healthcare literature explicitly addresses routing and scheduling of mobile service providers in order to treat patients whose conditions worsen over time. Several interesting parallels with our work can be found within this area. Nikzad et al. (2021) develop heuristic algorithms for stochastic home health care planning, with uncertain service times and travel times between patient locations included in their optimization framework. Cire and Diamant (2022) use approximate dynamic programming (ADP) methods to address the dynamic

allocation of home care patients to medical providers, assuming that patients are heterogeneous with respect to health requirements, service durations and regions of residence. Naderi et al. (2023) develop integrated approaches for staffing, assignment, routing and scheduling under uncertainty and propose a mixed-integer program (MIP) with a special structure that allows for a Benders decomposition. Nasir and Kuo (2024) consider stochastic models for home care transportation with dynamically prioritized patients and propose a chance-constrained multi-period optimization model, in which patient priorities are translated into time-dependent healthcare costs. Khorasanian et al. (2024) also study dynamic home care routing within an MDP framework and consider uncertainty in the number of visits per patient referral. Emergency department and hospital settings provide additional healthcare applications involving patient queues and resource allocation. Patrick et al. (2008) develop MDP models for the dynamic scheduling of patients with different priorities to diagnostic facilities, with an objective based on waiting time targets. Lee and Lee (2020) apply DRL to emergency department patient scheduling, using an MDP model that accounts for diverse patient acuity levels and personalized treatment plans. In the context of our paper, the relevant themes in the healthcare studies that we have cited include stochastically evolving demands across multiple facilities, limited service capacity, and the need to allocate resources dynamically in response to costs that become steeper as demand remains unresolved.

Infrastructure monitoring systems can be regarded as another important application domain. In these systems, critical assets are often classified into discrete condition grades based on inspection criteria, and limited maintenance resources must be allocated dynamically. Bridges, roads and other infrastructure naturally degrade over time due to traffic loads, environmental factors, and ageing, and this creates systems with natural finite-state representations and stochastic deterioration processes. Several relevant studies can be found in the recent literature. Tao et al. (2021) consider the maintenance of highway bridges and formulate a model that innovatively combines discrete-time and continuous-time MDPs (based on different possible causes of degradation) to find optimal maintenance policies. Xu and Guo (2022) also consider bridge maintenance and begin by using a Q-learning algorithm to optimize the maintenance strategy for a single bridge, then use integer programming to optimize budget allocation for a network of bridges. In related work, Xu and Azhari (2022) employ hazard-based approaches to optimize both inspection intervals and repair decisions via semi-Markov decision processes (SMDPs). Xin et al. (2023) also apply SMDPs to bridge element maintenance, using Weibull distributions for state duration modeling and choosing between maintenance strategies based on specific defect types and development states. Liu et al. (2025) have also studied dynamic optimization methods for bridge networks under budget constraints. Recent work by Fairley et al. (2025) investigates the optimal design of failure-prone systems, with first-stage strategic planning decisions used to decide the number of components (which could represent transport links, for example) to install in a stochastic system that will be maintained dynamically in the second stage.

In summary, this section has shown that problems involving dynamic allocation of limited resources to stochastically degrading systems can be found across diverse application domains.

Common modeling elements include stochastic arrival and service processes, finite-state representations of system conditions, cost functions that increase with unserved demand or degradation levels, and explicit consideration of travel or transition times between service locations.

### 3 Problem formulation

Let  $G = (V, E)$  be a connected graph, referred to as a *network*, where  $V$  and  $E$  are the sets of nodes and edges respectively. Let  $M \subseteq V$  denote a subset of nodes referred to as *machines*, and let  $N := V \setminus M$  denote the set of other nodes, referred to as *intermediate stages*. We use  $m := |M|$  (resp.  $n := |N|$ ) to denote the number of machines (resp. intermediate stages). As a convention we label the machines  $1, 2, \dots, m$ , while the stages in  $N$  are labeled  $m + 1, m + 2, \dots, m + n$ .

At any given time, machine  $i \in M$  is in state  $x_i \in \{0, 1, \dots, K_i\}$ , where 0 is the pristine or ‘as-good-as-new’ state and  $K_i$  is the ‘failed’ state. When machine  $i$  is in state  $x_i \leq K_i - 1$ , transitions to state  $x_i + 1$  occur at an exponential rate  $\lambda_i > 0$ ; this is the *degradation rate* for machine  $i$ . All machines in  $M$  are assumed to degrade independently of each other. In addition, a cost  $f_i(x_i)$  is incurred per unit time while machine  $i$  is in state  $x_i$ . We assume that the functions  $f_i$  are strictly increasing and that  $f_i(0) = 0$  for each  $i \in M$ .

There is a single repairer who occupies a single node in  $V$  at any given time. At each point in time, the repairer can decide either to remain at the currently-occupied node or attempt to move to one of the adjacent nodes (connected by edges to the current node). In the latter case, it must also choose which node to move to. Therefore, if the number of adjacent nodes is  $l$  then the number of possible decisions for the repairer is  $l + 1$ . If it remains at a node  $i \in M$  (i.e. at a machine) and  $x_i \geq 1$  then it is assumed to be *repairing* the machine, in which case the state of the machine transitions from  $x_i$  to  $x_i - 1$  at a rate  $\mu_i > 0$ , where  $\mu_i$  is the *repair rate* for machine  $i$ . We note that degradations can still happen at a machine while it’s being repaired, so that  $\lambda_i + \mu_i$  is the total transition rate for a non-failed machine  $i$  while it is under repair. If the repairer remains at a node  $i$  with  $x_i = 0$ , then the repairer is said to be *idle*. Similarly, idleness can also occur if the repairer chooses to remain at some stage  $j \in N$  rather than attempting to move. On the other hand, if the repairer chooses to move to an adjacent node, then it moves to that node at an exponential rate  $\tau > 0$ , referred to as the *switching rate*. Repair times and switching times are assumed to be independent of the degradation processes for machines in  $M$ . We assume that the switching rate is the same between any adjacent pair of nodes, so that the expected time to move from one node to another is proportional to the number of edges that must be traversed. This implies that the amount of time to switch from one machine to another is distributed as a sum of i.i.d. exponential random variables and therefore follows an Erlang distribution.

The above assumptions allow the system to be formulated as a finite-state, continuous-time Markov decision process (MDP). Herein, we will follow similar notational conventions to those in Tian and Shone (2026). Let  $S$  be the state space, given by

$$S := \{(i, (x_1, \dots, x_m)) \mid i \in V, x_i \in \{0, 1, \dots, K_i\} \text{ for } i \in M\},$$

where  $i$  indicates the node currently occupied by the repairer. As a notational convention, we will tend to use  $i$  for the repairer's current location throughout this paper, while  $j$  will be used more generally for nodes in  $V$ . Also,  $k$  will be used to represent the condition of an individual machine, which we will sometimes refer to as the state of the machine (not to be confused with the system state). Given that the repairer cannot be repairing and switching at the same time (and also cannot be carrying out more than one repair), the sum of the transition rates under any state cannot exceed  $\sum_{j \in M} \lambda_j + \max\{\mu_1, \dots, \mu_m, \tau\}$  and we can therefore use the technique of uniformization (Lippman (1975); Serfozo (1979)) to consider an equivalent discrete-time formulation which evolves in time steps of size  $\Delta := \left(\sum_{j \in M} \lambda_j + \max\{\mu_1, \dots, \mu_m, \tau\}\right)^{-1}$ .

For each state  $\mathbf{x} = (i, (x_1, \dots, x_m)) \in S$ , let  $R(\mathbf{x})$  denote the set of nodes adjacent to node  $i$  and let  $A_{\mathbf{x}} = \{i\} \cup R(\mathbf{x})$  denote the set of actions available under state  $\mathbf{x}$ . We can interpret action  $a \in A_{\mathbf{x}}$  as the node that the repairer attempts to move to next, given that it is currently in state  $\mathbf{x}$  (with  $a = i$  indicating that the repairer remains at the current node). Then, for two distinct states  $\mathbf{x} := (i, (x_1, \dots, x_m))$  and  $\mathbf{x}' := (i', (x'_1, \dots, x'_m))$ , the transition probability of moving from state  $\mathbf{x}$  to  $\mathbf{x}'$  in the uniformized MDP can be expressed as

$$p_{\mathbf{x}, \mathbf{x}'}(a) := \begin{cases} \lambda_j \Delta, & \text{if } i' = i, x'_j = x_j + 1 \text{ for some } j \in M \text{ and } x'_l = x_l \text{ for } l \neq j, \\ \mu_i \Delta, & \text{if } i' = i, x'_i = x_i - 1, x'_j = x_j \text{ for } j \neq i \text{ and } a = i, \\ \tau \Delta, & \text{if } i' \in R(\mathbf{x}), x'_j = x_j \text{ for all } j \in M \text{ and } a = i', \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The first line in (1) corresponds to a degradation event at machine  $j$ , the second line corresponds to a completed (partial) repair at the repairer's current node  $i$  and the third line represents switching from node  $i$  to  $i'$ . We also have  $p_{\mathbf{x}, \mathbf{x}}(a) = 1 - \sum_{\mathbf{y} \neq \mathbf{x}} p_{\mathbf{x}, \mathbf{y}}(a)$  as the probability of remaining in the same state. Since the units of time are arbitrary, in the rest of the paper we will assume  $\Delta = 1$  without loss of generality. This implies that we can use quantities such as  $\lambda_j, \mu_j, \tau$  to represent transition probabilities in the uniformized MDP.

For each time step that the system spends in state  $\mathbf{x}$  a cost  $c(\mathbf{x})$  is incurred, given by

$$c(\mathbf{x}) = \sum_{i \in M} f_i(x_i). \quad (2)$$

Let  $\theta$  denote the decision-making policy to be followed, i.e. the method of choosing actions in our MDP. If the same action  $\theta(\mathbf{x}) \in A_{\mathbf{x}}$  is chosen every time the system is in state  $\mathbf{x} \in S$ , then the policy is referred to as *stationary* and *deterministic* (Puterman (1994)). The expected long-run average cost (or *average cost* for short) under policy  $\theta$ , given that the system begins in state  $\mathbf{x}_0 \in S$ , is

$$g_{\theta}(\mathbf{x}) = \liminf_{t \rightarrow \infty} t^{-1} \mathbb{E}_{\theta} \left[ \sum_{n=0}^{t-1} c(\mathbf{x}_n) \mid \mathbf{x}_0 = \mathbf{x} \right], \quad (3)$$

where  $\mathbf{x}_n$  is the system state at time step  $n \in \mathbb{N}_0$ . The theory of uniformization implies that if  $\theta$  is a stationary policy then the average cost  $g_\theta(\mathbf{x})$  in the discretized system is equal to the average cost per unit time incurred under the same policy in the continuous-time system (in which we suppose that actions are chosen at the beginning of each new sojourn, i.e. every time the system transitions from one state to another). The objective of the problem is to find a policy  $\theta^*$  such that

$$g_{\theta^*}(\mathbf{x}) \leq g_\theta(\mathbf{x}) \quad \forall \theta \in \Theta, \mathbf{x} \in S,$$

where  $\Theta$  is the set of all admissible policies. In other words, we aim to find a policy that minimizes the average cost.

Since  $S$  is finite, the question of system stability does not arise. However, our MDP formulation does have some particular subtleties. For one thing, we cannot claim that the MDP is *unichain* (a common property of MDPs based on queueing systems, for example), since one can easily define a stationary policy  $\theta$  under which the resulting Markov chain has more than one positive recurrent class. The simplest way to do this is to define the actions  $\theta(\mathbf{x})$  in such a way that the repairer always remains at whichever node it currently occupies. This policy allows all machines except one (if the machine is initially located at a node in  $M$ ) to remain permanently in their failed states. The average cost under such a policy is finite, but clearly depends on the initial state. However, we can show that our MDP belongs to the ‘communicating’ class of multichain MDP models, as discussed in Puterman (1994) (p. 348), from which it follows that the average cost under an optimal policy must be independent of the initial state. We state this formally below and provide a proof in Appendix A.

**Proposition 3.1.** *Let  $\theta^*$  be an optimal policy. Then the average cost  $g_{\theta^*}(\mathbf{x})$  is independent of the initial state  $\mathbf{x}$ , and can be expressed as a constant  $g^*$ .*

The theory in (Puterman (1994)) then implies that there exists a function  $h : S \rightarrow \mathbb{R}$  satisfying the optimality equations

$$g^* + h(\mathbf{x}) = c(\mathbf{x}) + \min_{a \in A_{\mathbf{x}}} \left\{ \sum_{\mathbf{y} \in S} p_{\mathbf{x},\mathbf{y}}(a) h(\mathbf{y}) \right\} \quad \forall \mathbf{x} \in S. \quad (4)$$

In theory, we can use dynamic programming (DP) techniques to compute a constant  $g^*$  and function  $h$  satisfying the optimality equations (4). DP can be summarized as a framework for solving MDPs via the recursive solution of subproblems using Bellman’s optimality principle (Bellman (1954)), which essentially states that the optimal solution of a multi-stage problem is composed of optimal solutions to smaller subproblems. Common DP solution algorithms include value iteration, which iteratively updates estimates of the value function  $h$  using Bellman’s optimality equation, and policy iteration, which involves alternating between evaluating a fixed policy and improving it via a greedy improvement step. These algorithms are made more complicated by the existence of policies with multichain structure, but the ‘communicating’ property ensures their theoretical feasibility

(see Puterman (1994), pp. 478-484). Although DP algorithms are useful for finding optimal solutions in small problem instances, they become computationally intractable as  $S$  increases in size. We therefore propose the use of index heuristics and reinforcement learning in Sections 4 and 5 respectively in order to find strong-performing policies in systems with large state spaces.

Before closing this section, we present a short example to show that the finite-state nature of our problem prevents direct application of certain results that have been proved for infinite-state problems with similar formulations. As mentioned in our introduction, Duenyas and Van Oyen (1996) considers an infinite-state problem in which a server must be allocated dynamically between different queues. Their formulation is different from ours in several respects; for example, it assumes that switching and service times are non-interruptible. They also assume linear holding costs and, under this assumption, are able to show that different queues can be assigned priorities based on the values of  $c_i\mu_i$ , where  $c_i$  is the linear holding cost for customers at queue  $i$  and  $\mu_i$  is the service rate. More specifically, they show that any queue  $i$  belonging to the set  $\arg \max_i \{c_i\mu_i\}$  can be defined as a ‘top-priority’ queue and, under an optimal policy, the server should never switch away from such a queue if it is non-empty. The next example, which considers the case of linear holding costs, shows that the analogous property does not hold in our system.

**Example 3.2.** *Consider a network consisting of only 2 nodes connected by a single edge. Both nodes are machines; that is,  $M = \{1, 2\}$ . We assume that  $\lambda_1 = \lambda_2 = 0.4$ ,  $K_1 = K_2 = 2$  and  $f_1(x) = f_2(x) = x$  for  $x \in \{0, 1, 2\}$ ; that is, the degradation rates for both machines are the same, they both have 3 possible states (with state 2 being the ‘failed’ state) and they both have the same cost function, which is a linear function of the degree of degradation. The repair rates are  $\mu_1 = 1.1$  and  $\mu_2 = 1$ , so machine 1 can be repaired slightly faster than machine 2. The switch rate is  $\tau = 100$ . Table 1 shows an optimal policy for this system, computed using dynamic programming. To clarify, part (a) of Table 1 shows the decisions specified by the optimal policy at the 9 possible states where the repairer is at machine 1, and part (b) shows the decisions specified at the 9 possible states where the repairer is at machine 2.*

Table 1: Decisions made under an optimal policy for Example 3.2

(a) when the repairer is at machine 1				(b) when the repairer is at machine 2			
	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$		$x_2 = 0$	$x_2 = 1$	$x_2 = 2$
$x_1 = 0$	1	2	2	$x_1 = 0$	1	2	2
$x_1 = 1$	1	1	1	$x_1 = 1$	1	1	1
$x_1 = 2$	1	2	1	$x_1 = 2$	1	2	1

*Under the optimal policy, the decision under state  $\mathbf{x} = (1, (2, 1))$  is to switch to machine 2. Thus, the repairer moves away from machine 1 despite the fact that  $c_1\mu_1 > c_2\mu_2$  and machine 1 requires repair ( $x_1 > 0$ ). This shows that in a finite-state problem, the ‘top-priority’ rule described in (Duenyas and Van Oyen (1996)) no longer applies.*

## 4 Index heuristic

In order to develop an index-based heuristic for our dynamic repair and maintenance problem, we begin by proving that the cost function (2) is equivalent (in terms of average cost incurred under a fixed stationary policy) to an alternative function that focuses attention on the repairer's current location. For a general state  $\mathbf{x} = (i, (x_1, \dots, x_m)) \in S$  and action  $a \in A_{\mathbf{x}}$ , let  $\tilde{c}(\mathbf{x}, a)$  be defined as follows:

$$\tilde{c}(\mathbf{x}, a) = \begin{cases} \sum_{j \in M} f_j(K_j) - \frac{\mu_i}{\lambda_i} [f_i(K_i) - f_i(x_i - 1)], & \text{if } i \in M, x_i \geq 1 \text{ and } a = i, \\ \sum_{j \in M} f_j(K_j), & \text{otherwise.} \end{cases} \quad (5)$$

Notably, the function  $\tilde{c}$  (unlike the original cost function  $c$ ) depends on both the state  $\mathbf{x}$  and the action  $a$  chosen under  $\mathbf{x}$ . We observe that  $\tilde{c}(\mathbf{x}, a)$  includes a negative term if and only if the repairer is located at a damaged machine  $i$  and chooses to perform repairs. This negative term increases in magnitude as the machine approaches the pristine state. The next result establishes the equivalence between the cost functions  $c$  and  $\tilde{c}$  under stationary policies. We provide a proof in Appendix B.

**Proposition 4.1.** *Let  $\theta$  be a stationary policy, and let*

$$\tilde{g}_\theta(\mathbf{x}) := \liminf_{t \rightarrow \infty} t^{-1} \mathbb{E}_\theta \left[ \sum_{n=0}^{t-1} \tilde{c}(\mathbf{x}_n, \theta(\mathbf{x}_n)) \mid \mathbf{x}_0 = \mathbf{x} \right], \quad \mathbf{x} \in S, \quad (6)$$

where  $\mathbf{x}_n$  is the system state at time step  $n \in \mathbb{N}_0$ . That is,  $\tilde{g}_\theta(\mathbf{x})$  is the average cost under policy  $\theta$  given that we replace the cost function  $c(\mathbf{x})$  by  $\tilde{c}(\mathbf{x}, a)$ . Then

$$\tilde{g}_\theta(\mathbf{x}) = g_\theta(\mathbf{x}) \quad \forall \mathbf{x} \in S, \quad (7)$$

where  $g_\theta(\mathbf{x})$  is the average cost defined in (3).

The intuition for Proposition 4.1 is as follows. If the repairer adopts a completely passive policy and never repairs any machines, then the long-run average cost is trivially equal to  $\sum_{j \in M} f_j(K_j)$ . Suppose we modify the passive policy by choosing a specific state  $\mathbf{x} = (i, (x_1, \dots, x_m))$  with  $x_i \geq 1$ , and specifying that the repairer should always choose to repair machine  $i$  under this state. Given that the repairer chooses this action at a particular time step, there is a probability  $\mu_i$  that the machine's state transitions to  $x_i - 1$  at the next time step. Following this transition, we obtain a cost rate advantage of  $[f_i(x_i) - f_i(x_i - 1)]$  compared to the passive policy (under which the machine would have remained in state  $x_i$ ). This advantage persists until the machine next degrades, which is expected to happen  $1/\lambda_i$  time steps later. At that point, the cost rate advantage becomes  $[f_i(x_i + 1) - f_i(x_i)]$  and this also continues for an expected  $1/\lambda_i$  time steps. By considering a sequence of further degradations until machine  $i$  eventually degrades to state  $K_i$ , we can quantify the advantage of choosing the repair action under state  $\mathbf{x}$  as  $\mu_i [f_i(K_i) - f_i(x_i - 1)] / \lambda_i$ . It follows

that this quantity should be deducted from the average cost of the passive policy every time the repair action is chosen at  $\mathbf{x}$ . In Appendix B we give a proof using more rigorous arguments, by considering the detailed balance equations of the continuous-time Markov chain induced by a fixed stationary policy.

Proposition 4.1 enables us to consider an equivalent MDP formulation based on maximizing rewards, rather than minimizing costs. The reward under a particular state  $\mathbf{x} \in S$  is equal to the absolute value of the negative term in (5) if the repairer performs a repair under  $\mathbf{x}$ , and zero otherwise. The next result formalizes this equivalence.

**Corollary 4.2.** *Consider an MDP formulation in which the single-step reward for being in state  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and choosing action  $a \in A_{\mathbf{x}}$ , denoted by  $r(\mathbf{x}, a)$ , is given by*

$$r(\mathbf{x}, a) = \begin{cases} \frac{\mu_i}{\lambda_i} [f_i(K_i) - f_i(x_i - 1)], & \text{if } i \in M, x_i \geq 1 \text{ and } a = i, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

and the objective is to maximize the long-run average reward, defined as

$$u_{\theta}(\mathbf{x}) := \limsup_{t \rightarrow \infty} t^{-1} \mathbb{E}_{\theta} \left[ \sum_{n=0}^{t-1} r(\mathbf{x}_n, \theta(\mathbf{x}_n)) \mid \mathbf{x}_0 = \mathbf{x} \right], \quad \mathbf{x} \in S. \quad (9)$$

All other aspects of the MDP formulation, including states, actions and transition probabilities, are the same as described in Section 3. Then any stationary policy which maximizes the long-run average reward (9) in the new MDP formulation also minimizes the long-run average cost (3) in the previous MDP formulation.

*Proof.* By Proposition 4.1 it is sufficient to show that maximizing the long-run average reward (9) is equivalent to minimizing the modified average cost (6) in which the alternative cost function  $\tilde{c}(\mathbf{x}, a)$  is used. Indeed, the cost function  $\tilde{c}$  defined in (5) includes a constant term  $\sum_{j \in M} f_j(K_j)$  that is incurred under all states. It follows that  $\tilde{g}_{\theta}(\mathbf{x}) = \sum_{j \in M} f_j(K_j) - u_{\theta}(\mathbf{x})$ , and hence minimizing  $\tilde{g}_{\theta}(\mathbf{x})$  is equivalent to maximizing  $u_{\theta}(\mathbf{x})$ .  $\square$

Recall that, by Proposition 3.1, the optimal average cost  $g^*$  is independent of the initial state. Therefore Corollary 4.2 implies that the optimal average reward  $u^* = \sup_{\theta} u_{\theta}(\mathbf{x})$  also has no dependence on  $\mathbf{x}$ .

The main advantage of using the reward-based formulation (9) is that it allows the reward under a particular state  $\mathbf{x}$  to be computed purely in terms of the parameters for the repairer's current location  $i$  (with the reward being zero if  $i$  is an intermediate stage, rather than a machine), whereas the cost formulation (3) involves aggregating cost rates over all of the damaged machines under a particular state. This greatly assists the development of an index-based heuristic policy. A similar approach was used by (Duenyas and Van Oyen (1996)), who formulated a problem based on minimizing average costs but then developed an index heuristic based on maximizing rewards.

In the infinite-state problem considered by (Duenyas and Van Oyen (1996)), the reward function takes the much simpler form  $c_i\mu_i$  (where, in their notation,  $c_i$  is the linear cost rate associated with queue  $i$ ). The reward function (9) in our model is more complicated as it depends on the level of degradation,  $x_i$ , and also on the degradation rate  $\lambda_i$ .

We now proceed to describe the development of an index-based heuristic for our MDP, focusing on the single-step reward function  $r(\mathbf{x}, a)$  rather than the cost function  $c(\mathbf{x})$  from Section 3. The basic principle underlying the heuristic is that, under any given state  $\mathbf{x} \in S$ , the repairer calculates ‘average reward rates’ (referred to as *indices*) associated with different possible actions, and then chooses an action that maximizes the index. These indices are calculated over a short time horizon, by assuming that the repairer will commit to fully repairing the next machine that it visits. More specifically, if the repairer is located at a machine  $i \in M$  then the index for remaining at node  $i$  is calculated by assuming that the machine remains there until  $x_i = 0$ , whereas the index for switching to an alternative node  $j \neq i$  is calculated by assuming that the machine moves directly to node  $j$  and then remains there until  $x_j = 0$ . If the repairer is at an intermediate stage  $i \in N$  then the procedure is similar, except we only calculate indices for moving to machines in  $j \in M$  and do not calculate an index for remaining at  $i$ . These indices inform the action chosen by the repairer at the current time step, but the indices are recalculated at every time step, so (for example) the repairer might begin moving towards a particular machine  $j$  but change direction and move towards a different machine  $l$  if the latter machine experiences a degradation event.

Firstly, for a particular machine  $i \in M$ , let  $T_i(k)$  denote the random amount of time taken for the machine to return to its pristine state, given that it begins in state  $k \in \{0, 1, \dots, K_i\}$  and the repairer begins to repair it immediately and continues without any interruption. Also, let  $R_i(k)$  denote the cumulative reward earned while these repairs are in progress, i.e. until the machine returns to the pristine state. We define both  $T_i(0)$  and  $R_i(0)$  to be zero. Expressions for  $\mathbb{E}[T_i(k)]$  and  $\mathbb{E}[R_i(k)]$  for  $k \geq 1$  can be derived quite easily in terms of the system parameters, by considering the dynamics of an  $M/M/1$  queue with finite capacity. These expressions are given in the next lemma.

**Lemma 4.3.** *For each machine  $i \in M$  and  $k \in \{1, \dots, K_i\}$ , we have*

$$\mathbb{E}[R_i(k)] = \sum_{p=1}^{K_i} C_{ip}(k) s_i(k), \quad \mathbb{E}[T_i(k)] = \sum_{p=1}^{K_i} C_{ip}(k), \quad (10)$$

where  $s_i(k) := \mu_i[f_i(K_i) - f_i(k-1)]/\lambda_i$  is the reward rate in state  $k$  and the coefficients  $C_{ip}(k)$  are given by

$$C_{ip}(k) = \begin{cases} \frac{1}{\mu_i^p} \sum_{r=0}^{p-1} \lambda_i^{p-1-r} \mu_i^r, & p = 1, 2, \dots, k, \\ \frac{1}{\mu_i^p} \sum_{r=0}^{k-1} \lambda_i^{p-1-r} \mu_i^r, & p = k+1, k+2, \dots, K_i. \end{cases}$$

*Proof.* Given that the repairer performs uninterrupted repairs at machine  $i$ , standard arguments based on the dynamics of continuous-time Markov chains imply that

$$\mathbb{E}[R_i(k)] = \begin{cases} \frac{s_i(k)}{\lambda_i + \mu_i} + \frac{\lambda_i}{\lambda_i + \mu_i} \mathbb{E}[R_i(k+1)] + \frac{\mu_i}{\lambda_i + \mu_i} \mathbb{E}[R_i(k-1)], & k = 1, 2, \dots, K_i - 1, \\ \frac{s_i(K_i)}{\mu_i} + \mathbb{E}[R_i(K_i - 1)], & k = K_i. \end{cases} \quad (11)$$

Thus, we have a system of  $K_i$  equations in  $K_i$  unknowns and can solve these to find the values of  $\mathbb{E}[R_i(1)], \dots, \mathbb{E}[R_i(K_i)]$  stated in (10).

For the  $\mathbb{E}[T_i(k)]$  values, the equations (11) take exactly the same form except that we replace each  $R_i(k)$  by  $T_i(k)$  and each reward rate  $s_i(k)$  should be replaced by 1. The equations can then be solved to give the results stated in (10).  $\square$

Suppose the system is in state  $\mathbf{x} = (i, (x_1, \dots, x_m))$ , where  $i \in M$ . We define the index for remaining at machine  $i$ , denoted by  $\Phi_i^{\text{stay}}(x_i)$ , as

$$\Phi_i^{\text{stay}}(x_i) = \begin{cases} \frac{\mathbb{E}[R_i(x_i)]}{\mathbb{E}[T_i(x_i)]}, & \text{if } x_i \in \{1, \dots, K_i\}, \\ 0, & \text{if } x_i = 0. \end{cases} \quad (12)$$

Next, let the repairer's current location  $i$  be an arbitrary node in  $V$ , and let  $D_{ij}$  denote the random amount of time needed for the repairer to move from node  $i$  to a machine  $j \neq i$ , assuming that it follows a shortest path in the network. We define the index for moving to machine  $j$ , denoted by  $\Phi_{ij}^{\text{move}}(x_j)$ , as

$$\Phi_{ij}^{\text{move}}(x_j) := \sum_{k=x_j}^{K_j} \mathbb{P}(X_j = k) \frac{\mathbb{E}[R_j(k)]}{\mathbb{E}[D_{ij}|X_j = k] + \mathbb{E}[T_j(k)]}, \quad (13)$$

where  $X_j$  is the (random) state of machine  $j$  upon arrival of the repairer at node  $j$ , suppressing the dependence on  $x_j$  for convenience. We note that  $X_j$  is distributed as a sum of i.i.d. geometric random variables and therefore follows a truncated version of the negative binomial distribution, as follows:

$$\mathbb{P}(X_j = k) = \begin{cases} \binom{d_{ij} + k - x_j - 1}{d_{ij} - 1} \left( \frac{\tau}{\lambda_j + \tau} \right)^{d_{ij}} \left( \frac{\lambda_j}{\lambda_j + \tau} \right)^{k - x_j}, & k = x_j, \dots, K_j - 1, \\ 1 - \sum_{r=x_j}^{K_j-1} \mathbb{P}(X_j = r), & k = K_j, \end{cases}$$

where  $d_{ij}$  is the number of nodes that the repairer must traverse on a shortest path from  $i$  to  $j$ .

For  $k = x_j, \dots, K_j - 1$  we have

$$\mathbb{E}[D_{ij}|X_j = k] = \frac{d_{ij} + k - x_j}{\tau + \lambda_j},$$

while for  $k = K_j$  we use the law of total expectation, which implies that

$$\mathbb{E}[D_{ij}] = \frac{d_{ij}}{\tau} = \sum_{k=x_j}^{K_j} \mathbb{P}(X_j = k) \mathbb{E}[D_{ij}|X_j = k],$$

and hence

$$\mathbb{E}[D_{ij}|X_j = K_j] = \frac{(d_{ij}/\tau) - \sum_{k=x_j}^{K_j-1} \mathbb{P}(X_j = k) \mathbb{E}[D_{ij}|X_j = k]}{1 - \sum_{k=x_j}^{K_j-1} \mathbb{P}(X_j = k)}.$$

We may interpret the index  $\Phi_i^{\text{stay}}(x_i)$  as an approximate measure of the average reward per unit time that the repairer can obtain by remaining at machine  $i$  until  $x_i = 0$ , while  $\Phi_{ij}^{\text{move}}(x_j)$  is the approximate average reward per unit time for switching to machine  $j$  and remaining there until  $x_j = 0$ . These are only approximate measures because, in fact, random variables such as  $R_j(k)$  and  $T_j(k)$  are not independent of each other and so the quotients that appear in (12) and (13) cannot be regarded as exact closed forms for the average rewards that would be obtained during the relevant time periods.

There is one further index quantity that should be defined. Suppose the system is in state  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and let  $\Phi_{ij}^{\text{wait}}(x_j)$  be defined in a similar way to  $\Phi_{ij}^{\text{move}}(x_j)$  except that it represents the (approximate) average reward that the repairer would gain by waiting for one further degradation event to occur at machine  $j \neq i$ , then carrying out the same actions as in the case of the previous index (that is, switching to  $j$  and repairing it until  $x_j = 0$ ). If  $x_j = K_j$  then no further degradations at  $j$  are possible, but we still suppose that the repairer waits for  $1/\lambda_j$  time units before moving to  $j$ . Specifically, we define

$$\begin{aligned} \Phi_{ij}^{\text{wait}}(x_j) := & \sum_{k=x_j}^{K_j-1} \mathbb{P}(X_j = k) \frac{\mathbb{E}[R_j(k+1)]}{1/\lambda_j + \mathbb{E}[D_{ij}|X_j = k] + \mathbb{E}[T_j(k+1)]}, \\ & + \mathbb{P}(X_j = K_j) \frac{\mathbb{E}[R_j(K_j)]}{1/\lambda_j + \mathbb{E}[D_{ij}|X_j = K_j] + \mathbb{E}[T_j(K_j)]}. \end{aligned} \quad (14)$$

Note that in (14),  $X_j$  has the same definition as before (i.e. the state of machine  $j$  when the repairer arrives, given that the repairer begins moving immediately), but we calculate the index using  $R_j(k+1)$  and  $T_j(k+1)$  rather than  $R_j(k)$  and  $T_j(k)$  because the repairer waits for an extra degradation event before beginning to move. If  $\Phi_{ij}^{\text{move}}(x_j) < \Phi_{ij}^{\text{wait}}(x_j)$  then we can interpret this as an incentive for the repairer to wait longer before moving to node  $j$ , since the occurrence of a further degradation event will increase the average reward that can be earned by switching to  $j$  and repairing the machine fully. Note that if  $x_j = K_j$  then it is guaranteed that  $\Phi_{ij}^{\text{move}}(x_j) > \Phi_{ij}^{\text{wait}}(x_j)$ , so there is no incentive to wait in that case. In our index heuristic, we prevent the repairer from

switching from machine  $i$  to machine  $j$  if  $\Phi_{ij}^{\text{move}}(x_j) < \Phi_{ij}^{\text{wait}}(x_j)$ .

Before defining the index heuristic, we discuss how it should make decisions under states with  $x_j = 0$  for all  $j \in M$ . States of this form are quite different from others in  $S$ , as the repairer does not have any work to do. However, one might suppose that the repairer should still take some kind of preparatory action in order to ensure that it is able to respond to any future jobs in the most efficient manner. Accordingly, we aim to identify a particular node in the network that can be regarded as an ‘optimal position’ for the repairer while it is waiting for new tasks to materialize. Specifically, for each node  $i \in V$ , let  $\Psi(i)$  be defined as follows:

$$\Psi(i) := \sum_{j \in M} \left( \frac{\lambda_j}{\lambda_1 + \dots + \lambda_m} \right) \mathbb{E}[D_{ij}], \quad (15)$$

Thus,  $\Psi(i)$  is the expected amount of time spent moving to the next machine that degrades, where the expectation is taken with respect to which machine this will be. We define the repairer’s optimal idling position as the node  $i \in V$  that minimizes  $\Psi(i)$ . If there are multiple such nodes, then an arbitrary choice of idling position can be made.

The next definition specifies how actions are chosen by the index heuristic.

**Definition 4.4.** (*Index policy.*) Let  $\theta^{IND}$  be a stationary policy that uses the following rules to select an action under state  $\mathbf{x} = (i, (x_1, \dots, x_m)) \in S$ :

1. If  $x_j = 0$  for all  $j \in M$  (that is, all machines are in the pristine condition) then let  $i^* \in \arg \min_{i \in V} \Psi(i)$ . The repairer’s action should be to move to the first node on the shortest path from  $i$  to  $i^*$  (if  $i \neq i^*$ ), or to remain at  $i^*$  (if  $i = i^*$ ).
2. If  $x_j \geq 1$  for at least one  $j \in M$ , then consider two possible cases:
  - (a) If  $i \in M$  (that is, the repairer is at a machine) then let the set  $J$  be defined by

$$J := \{j \in M \setminus \{i\} : \Phi_{ij}^{\text{move}}(x_j) \geq \Phi_{ij}^{\text{wait}}(x_j)\}. \quad (16)$$

If  $J$  is non-empty, then let  $j^*$  be the machine in  $J$  that maximizes  $\Phi_{ij}^{\text{move}}(x_j)$ . If  $\Phi_{ij^*}^{\text{move}}(x_{j^*}) > \Phi_i^{\text{stay}}(x_i)$  then the repairer’s action should be to move to the first node on the shortest path from  $i$  to  $j^*$ . On the other hand, if either  $J$  is empty or  $\Phi_{ij^*}^{\text{move}}(x_{j^*}) \leq \Phi_i^{\text{stay}}(x_i)$ , then the repairer’s action should be to remain at node  $i$ .

- (b) If  $i \in N$  (that is, the repairer is at an intermediate stage), then let  $j^*$  be the machine that maximizes  $\Phi_{ij}^{\text{move}}(x_j)$ . The repairer’s action should be to move to the first node on the shortest path from  $i$  to  $j^*$ .

In cases where two or more actions are viable choices according to the above rules, we assume that the action chosen under  $\theta^{IND}$  is based on a pre-determined priority ordering of the nodes in  $V$ . We refer to  $\theta^{IND}$  as the index heuristic policy.

It is clearly of interest to investigate how well the index heuristic policy performs in comparison to an optimal policy, and also how well it performs relative to other heuristic policies. In Section 6 we provide results from numerical experiments in order to demonstrate its performance empirically. Theoretical performance guarantees (e.g. suboptimality bounds) are difficult to establish in the full generality of our model. However, we are able to show that the index policy is optimal in certain special cases of our problem. These special cases are discussed in the next two propositions.

**Proposition 4.5.** *Suppose the following 3 conditions are satisfied:*

- (i) *All machines are directly connected to each other; that is,  $V$  is a complete graph.*
- (ii) *Each machine has only two possible states; that is,  $K_j = 1$  for each  $j \in M$ .*
- (iii) *The degradation rates, repair rates and cost functions are the same for all machines. That is, we have  $\lambda_1 = \dots = \lambda_m$ ,  $\mu_1 = \dots = \mu_m$  and  $f_1(1) = \dots = f_m(1)$ .*

*Then the index policy is optimal.*

The proof of Proposition 4.5 can be found in Appendix C. This proposition describes a homogeneous type of problem in which all machines have the same parameter values, and furthermore assumes that the state of a machine is a binary variable, so that it is effectively either ‘working’ or ‘not working’. In this type of binary model, the cost function  $f_j(k)$  for machine  $j \in M$  is determined by a single positive number,  $f_j(1)$ . While binary models such as these might appear quite simplistic, they are still potentially relevant to real-world applications and have been studied in the field of reliability modeling (see Lagos and Romero (2020); Cancela et al. (2022)), and we include them in our numerical experiments in Section 6. Another important condition in Proposition 4.5 is that all nodes are connected to each other, so that we can assume there are no intermediate stages (as if the network has intermediate stages, these will clearly never be visited under an optimal policy).

We have been able to confirm using experiments that all three conditions in Proposition 4.5 are necessary; that is, if any of them are omitted then the index policy may not be optimal. Details of these experiments can be found in Appendix D. However, if condition (i) is relaxed (that is, machines are no longer directly connected to each other) then the index policy may still be optimal, subject to another condition on the system parameters. To illustrate this we introduce another type of network, referred to as a ‘star network’, which we define below.

**Definition 4.6.** *Suppose there exists an intermediate stage  $s \in N$  such that, for any two distinct machines  $i, j \in M$ , the unique shortest path from  $i$  to  $j$  is a path of length  $2r$  (where  $r \in \mathbb{N}$ ) that passes through node  $s$ . Then the network  $V$  is called a star network, with  $s$  and  $r$  referred to as the center and radius respectively.*

In a star network, all machines are equidistant from each other (see Figure 2). The next proposition shows that the index policy is optimal in a star network under certain conditions, including two of the conditions from Proposition 4.5.

**Proposition 4.7.** *Let  $V$  be a star network with radius  $r \geq 1$ , and suppose that conditions (ii) and (iii) from Proposition 4.5 hold. Then the index policy is optimal, provided that*

$$\tau > 2r\lambda, \tag{17}$$

where  $\lambda$  is the common degradation rate for all machines.

Proof of Proposition 4.7 can be found in Appendix E. We note that the condition in 17 is equivalent to  $2r/\tau < 1/\lambda$ , which states that the expected amount of time for the repairer to move from one machine to another is smaller than the expected amount of time for a degradation event to occur at a non-degraded machine. This condition should often be satisfied in applications where failures or degradation events occur at a relatively slow rate, so that the repairer can frequently change location without additional machines degrading in the meantime. For example, this situation could arise in computing infrastructures where hardware faults occur on a much slower timescale than communication or switching delays. Under the stated conditions, the index policy is a ‘greedy’ policy that always prioritizes the nearest degraded machine (or moves towards the center of the network if there are no degraded machines), and we can show that such a policy is also optimal. Please see Appendix E for full details of the proof.

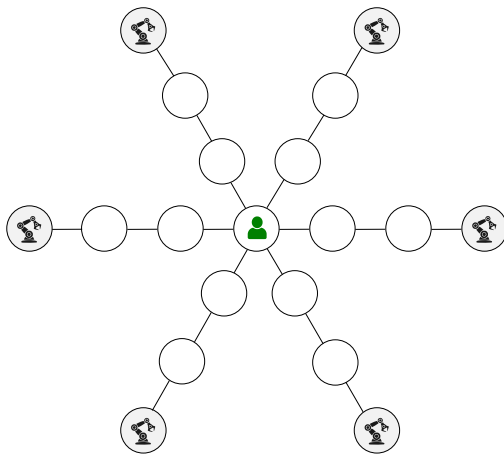


Figure 2: A star network with 6 machines and a radius  $r = 3$ .

## 5 Online policy improvement

The index heuristic developed in Section 4 yields a dynamic policy that, at any given time, either chooses a machine to prioritize next (if at least one machine requires repair), or directs the repairer to an optimal idling position (if no machines require repair). Importantly, in the former case, the indices in (12)-(14) are recalculated at each time step, so that the repairer can take advantage of the ability to interrupt switching and repair times as permitted under our problem formulation. However, a weakness of the index heuristic is that it works in a ‘short-sighted’ (myopic) way, as

the decision to prioritize a particular machine is made without taking into account that machine's proximity to other machines in the network. It would be preferable to use a more sophisticated approach that takes the network topology into account when making decisions.

As discussed in Section 3, an optimal policy can be characterized by a set of values  $\{h(\mathbf{x})\}_{\mathbf{x} \in S}$  satisfying the optimality equations in (4). The well-known method of *policy improvement* (PI), applicable to finite-state MDPs under the average cost criterion, works by starting from an arbitrary base policy and performing successive improvement steps, yielding a sequence of policies that eventually converges to an optimal one (Puterman (1994)). An optimal policy will naturally make long-sighted decisions that depend on the network topology, and any step of PI is therefore helpful in obtaining a policy that avoids short-sighted behavior. Although multiple improvement steps may be needed in order to obtain an optimal policy, it has been observed that the first improvement step often yields the largest improvement (Tijms (2003)). Furthermore, heuristics based on applying one step of policy improvement have been successfully applied to various types of queueing control problems (Krishnan (1990); Argon et al. (2009); Shone et al. (2020)).

As stated in Proposition 3.1, an optimal policy for our system is always unichain. However, the policy given by the index heuristic in Section 4 is not guaranteed to be unichain and indeed we have found that it can have a multichain structure in some problem instances. It will therefore be desirable to introduce a modified version of the index heuristic in which the resulting policy must be unichain, in order to avoid certain complications that arise when applying PI to a multichain base policy. The modified index heuristic is defined below.

**Definition 5.1.** (*Modified index policy.*) Let  $\theta^{\text{M-IND}}$  denote a stationary policy that uses the following rules to select an action under state  $\mathbf{x} = (i, (x_1, \dots, x_m)) \in S$ :

1. If  $x_j = K_j$  for all  $j \in M$  (that is, all machines are at their maximum state of degradation), then let  $j^* \in M$  be the smallest-indexed machine satisfying

$$j^* \in \arg \max_{j \in M} \left\{ \frac{\mathbb{E}[R_j(K_j)]}{\mathbb{E}[T_j(K_j)]} \right\}. \quad (18)$$

Then the action chosen under state  $\mathbf{x}$  should either be to move to the first node on the shortest path from  $i$  to  $j^*$  (if  $i \neq j^*$ ), or to stay at node  $i$  (if  $i = j^*$ ).

2. Otherwise, the action chosen under state  $\mathbf{x}$  should be the same action given by the unmodified index policy  $\theta^{\text{IND}}$  in Definition 4.4.

Note that under any admissible policy, it is always possible for the system to reach a state with  $x_j = K_j$  for all  $j \in M$ , since this can occur randomly via an unbroken sequence of machine degradations. Under the modified index policy  $\theta^{\text{M-IND}}$ , the machine  $j^*$  satisfying (18) is unique and has no dependence on the system state  $\mathbf{x}$ . Thus, the state  $(j^*, (K_1, \dots, K_m))$  is accessible from any state under  $\theta^{\text{M-IND}}$ . This ensures that there cannot be more than one positive recurrent class of states under  $\theta^{\text{M-IND}}$ , and therefore the policy is indeed unichain.

In general, the performance of  $\theta^{\text{M-IND}}$  may be worse than that of  $\theta^{\text{IND}}$ , since the rule that it uses under states with  $x_j = K_j$  for all  $j \in M$  is not consistent with the rules used under other states. The only purpose of introducing  $\theta^{\text{M-IND}}$  is so that it can be used as a convenient base policy in our PI method. In the computational experiments that we report in Section 6 we use the original definition of the index policy (Definition 4.4) when evaluating the performance of the index heuristic, and compare its performance with that of the simulation-based PI method.

In this section we propose an approximate one-step PI method, referred to as *online policy improvement* (OPI) for short, that works by simulating the random evolution of the system in the style of a reinforcement learning (RL) algorithm. RL refers to a broad class of methods in which decision-makers adaptively learn and improve their policies by interacting with a real or simulated environment, gaining experience by observing the costs or rewards from choosing different actions at the states that they encounter (Sutton and Barto (2018); Powell (2022)). Our OPI approach is divided into two stages, which we call the *offline* part and the *online* part. In the offline part, we repeatedly simulate the evolution of the system under the modified index policy  $\theta^{\text{M-IND}}$  in order to estimate the value function  $h_{\theta^{\text{M-IND}}}(\mathbf{x})$  under this policy. This closely resembles the method of *temporal difference learning* (TD-learning), discussed extensively in classical RL texts (see Sutton and Barto (2018), pp. 119-138), although our method modifies the classical approach by storing value function estimates only for a limited number of ‘exemplar’ states and allowing each simulation to continue until it reaches one of these exemplar states. Subsequently, in the online part, we simulate the evolution of the system and, at each discrete time step, choose an action by applying an improvement step to the base policy  $\theta^{\text{M-IND}}$ . Actions must be selected at the beginning of each time step, but we assume that after an action is selected, the next transition of the system does not occur until a further  $\delta$  time units have elapsed (here,  $\delta$  represents the length of a time step and is analogous to the uniformization parameter  $\Delta$  used in Section 3). During this  $\delta$ -length time period we run nested simulation experiments (within the main simulation) with the aim of improving the estimates of  $h_{\theta^{\text{M-IND}}}(\mathbf{x}')$  for states  $\mathbf{x}'$  that are near the current state  $\mathbf{x}$ . The results of the nested online simulations supplement the results of the offline simulations in order to estimate the value function as accurately as possible within a neighborhood of the current state. Once the time limit  $\delta$  is reached, we use a criterion based on a confidence interval to determine whether to select the best ‘improving’ action identified during the  $\delta$ -length period, or whether to select the action prescribed by the base policy (the latter option is chosen in situations where we cannot say with sufficient confidence which action is best).

The method that we use to find improving actions in the online part may be recognized as a type of *rollout* approach, as it involves comparing different actions under the current state by simulating future trajectories following each possible action, assuming that a fixed base policy will be used to choose actions at future time steps. Each simulated trajectory ends when the system reaches a state for which a value function estimate is stored, and the results from all simulated trajectories are aggregated to obtain estimates of the relative future costs under all possible actions. Rollout and direct lookahead methods have been extensively discussed in the recent RL literature (Bertsekas

(2020); Powell (2022) and offer the ability to make locally optimal decisions via simulation, thereby avoiding the modeling and computational challenges of other RL methods that involve constructing approximate value functions over the entire state space.

Some additional explanations of how the offline and online parts work are provided in the subsections below. To simplify notation we use  $\theta$  rather than  $\theta^{\text{M-IND}}$  to denote the modified index policy, referred to as the base policy. Due to space restrictions, we defer full details of the algorithms (including pseudocode) to Appendix F.

## 5.1 Offline part

In the offline part we simulate the evolution of the discrete-time MDP formulated in Section 3. In the preparatory stage, we conduct an initial simulation in order to identify a subset of states in  $Z \subset S$  that are visited often under the base policy  $\theta$ , and also obtain an estimate  $\hat{g}_\theta$  of the average cost under this policy. We then initialize an array  $U$ , used to store information about states visited during the simulation. Initially, the array  $U$  contains a single state which we denote by  $\mathbf{u}_0$ .

Next, we conduct a large number of variable-length simulations, where each simulation begins from one of the states in  $Z$  and continues until the system reaches one of the states in  $U$ . In general, suppose the  $r^{\text{th}}$  variable-length simulation (for  $r = 1, 2, \dots$ ) begins from state  $\mathbf{z}_r \in Z$  and ends in state  $\mathbf{u}_r \in U$ . Also, let  $\mathbf{x}$  be one of the first  $p$  states visited during this simulation, where  $p$  is a small integer (we use  $p = 5$  in our experiments). Then we calculate an estimate of  $h_\theta(\mathbf{x})$ , denoted by  $\hat{h}_\theta^{(s)}(\mathbf{x})$  (where  $s$  is a counter of how many estimates have been obtained for  $h_\theta(\mathbf{x})$  so far), as follows:

$$\hat{h}_\theta^{(s)}(\mathbf{x}) = C_r(\mathbf{z}_r, \mathbf{u}_r) + \hat{h}_\theta(\mathbf{u}_r) - \hat{g}_\theta T_r(\mathbf{z}_r, \mathbf{u}_r), \quad (19)$$

where  $C_r(\mathbf{z}_r, \mathbf{u}_r)$  is the total cost incurred during the variable-length simulation,  $T_r(\mathbf{z}_r, \mathbf{u}_r)$  is the total number of time steps in the simulation, and  $\hat{h}_\theta(\mathbf{u}_r)$  is the latest available estimate of  $h_\theta(\mathbf{u}_r)$  which is calculated using all of the estimates  $h_\theta^{(1)}(\mathbf{u}_r), h_\theta^{(2)}(\mathbf{u}_r)$ , etc. Note that equation (19) is derived using the well-known policy evaluation equations in dynamic programming, which are similar to the optimality equations (4). It can be explained intuitively as follows: in order to obtain an estimate of the relative value of being in state  $\mathbf{x}$ , we take the total cost incurred from starting in that state, which is  $C_r(\mathbf{z}_r, \mathbf{u}_r) + \hat{h}_\theta(\mathbf{u}_r)$  (where  $\hat{h}_\theta(\mathbf{u}_r)$  represents the ‘terminal cost’ associated with finishing in state  $\mathbf{u}_r$ ) and subtract the cost that would have been earned if we had incurred the long-run average cost  $\hat{g}_\theta$  over  $T_r(\mathbf{z}_r, \mathbf{u}_r)$  time steps.

After obtaining the new estimate  $\hat{h}_\theta^{(s)}(\mathbf{x})$  we add state  $\mathbf{x}$  to the array  $U$  if it is not included already. We also store some other information about  $\mathbf{x}$  in  $U$ , including the sum of the squared estimates  $[\hat{h}_\theta^{(s)}(\mathbf{x})]^2$ , which can be used to calculate a variance; this is needed in the online part of the algorithm (please see Appendix F for full details). As the offline part of the algorithm continues, the array  $U$  becomes populated with more states and more information about those states. After a large number of iterations, the offline part ends. The intended outcome is that  $U$  should include a subset of states that are visited often under the base policy and, for each state  $\mathbf{x} \in U$ , an estimate  $\hat{h}_\theta(\mathbf{x})$  which should be a reasonably accurate estimate of the true value  $h_\theta(\mathbf{x})$ . We have been able

to verify using experiments that the algorithm is able to output estimates  $\hat{h}_\theta(\mathbf{x})$  that are very close to the true values  $h_\theta(\mathbf{x})$  given by dynamic programming.

## 5.2 Online part

In this part we take the array  $U$  obtained from the offline part as an input and simulate the evolution of the system again, but instead of using the actions given by the base policy  $\theta$ , we apply an approximate policy improvement step. This means that if  $\mathbf{x}$  is the current state, we aim to select the action  $a$  that minimizes

$$\sum_{\mathbf{y} \in S} p_{\mathbf{x}, \mathbf{y}}(a) \hat{h}_\theta(\mathbf{y}), \quad (20)$$

where  $\hat{h}_\theta(\mathbf{y})$  is the latest available estimate of  $h_\theta(\mathbf{y})$ . This implies that if  $\mathbf{x} = (i, (x_1, \dots, x_m))$  denotes the current state then we require estimates  $\hat{h}_\theta(\mathbf{y})$  for all possible ‘neighboring’ states, i.e. all states  $\mathbf{y}$  such that  $p_{\mathbf{x}, \mathbf{y}}(a) > 0$  for some  $a \in A_{\mathbf{x}}$ . In general, let  $\mathbf{x}^{j+}$  (resp.  $\mathbf{x}^{j-}$ ) denote a state identical to  $\mathbf{x}$  except that the component  $x_j$  is increased (resp. decreased) by one, and let  $\mathbf{x}^{-j}$  denote a state identical to  $\mathbf{x}$  except that the repairer’s location is changed to  $j$ , where  $j \neq i$ . We also write  $\mathbf{x}^{-i} := \mathbf{x}$ . Note that degradation events occur independently of the action chosen, so any state of the form  $\mathbf{x}^{j+}$  for some  $j \in M$  can be ignored when it comes to minimizing (20), because it occurs with the same probability under any action. Therefore, to find the action that minimizes (20), we only need estimates  $\hat{h}_\theta(\mathbf{y})$  for states  $\mathbf{y}$  that are either (i) equal to  $\mathbf{x}$ , (ii) of the form  $\mathbf{y} = \mathbf{x}^{-j}$  where  $j$  is a node adjacent to  $i$ , or (iii) of the form  $\mathbf{x}^{i-}$  where  $i \in M$  is the repairer’s current location and  $x_i \geq 1$ . Let  $F(\mathbf{x})$  denote the set of states satisfying either (i), (ii) or (iii).

Immediately after the system enters state  $\mathbf{x}$ , we identify the action that minimizes (20) by using the latest available estimates  $\hat{h}_\theta(\mathbf{y})$  for states  $\mathbf{y} \in F(\mathbf{x})$ . However, rather than simply selecting the action that minimizes this expression, instead we use the information in  $U$  to calculate a confidence interval of the form  $[\hat{h}_\theta^-(\mathbf{y}), \hat{h}_\theta^+(\mathbf{y})]$  for each  $\mathbf{y} \in F(\mathbf{x})$ . We then say that an action  $a_1 \in A_{\mathbf{x}}$  is better than another action  $a_2 \in A_{\mathbf{x}}$  if and only if

$$\sum_{\mathbf{y} \in S} p_{\mathbf{x}, \mathbf{y}}(a_1) \hat{h}_\theta(\mathbf{y}) < \sum_{\mathbf{y} \in S} p_{\mathbf{x}, \mathbf{y}}(a_2) \hat{h}_\theta(\mathbf{y}) \quad \forall \hat{h}_\theta(\mathbf{y}) \in [\hat{h}_\theta^-(\mathbf{y}), \hat{h}_\theta^+(\mathbf{y})], \mathbf{y} \in F(\mathbf{x}). \quad (21)$$

In practice, (21) yields some simplifications. For example, if the repairer is at an intermediate stage of the network ( $i \notin M$ ) then the only possible actions are to switch to an adjacent node or to idle at the current node. After substituting the relevant transition probabilities into (21), we find that  $a_1$  should be preferred to  $a_2$  if

$$\hat{h}_\theta^+(\mathbf{x}^{-a_1}) < \hat{h}_\theta^-(\mathbf{x}^{-a_2}),$$

which means that the confidence interval for  $h_\theta(\mathbf{x}^{-a_1})$  is non-overlapping with that of  $h_\theta(\mathbf{x}^{-a_2})$  and lies entirely below it. On the other hand, if the repairer is at a machine  $i \in M$  with  $x_i \geq 1$  then we also need to consider whether remaining at node  $i$  (i.e. carrying out repairs) should be preferred to switching to  $j \neq i$ . In this case, remaining at  $i$  is better than switching to  $j$  if the

inequality

$$\mu_i \left( \hat{h}_\theta(\mathbf{x}^{i-}) - \hat{h}_\theta(\mathbf{x}) \right) < \tau \left( \hat{h}_\theta(\mathbf{x}^{\rightarrow j}) - \hat{h}_\theta(\mathbf{x}) \right)$$

holds for all  $\hat{h}_\theta(\mathbf{x}^{i-}) \in [\hat{h}_\theta^-(\mathbf{x}^{i-}), \hat{h}_\theta^+(\mathbf{x}^{i-})]$ ,  $\hat{h}_\theta(\mathbf{x}^{\rightarrow j}) \in [\hat{h}_\theta^-(\mathbf{x}^{\rightarrow j}), \hat{h}_\theta^+(\mathbf{x}^{\rightarrow j})]$  and  $\hat{h}_\theta(\mathbf{x}) \in [\hat{h}_\theta^-(\mathbf{x}), \hat{h}_\theta^+(\mathbf{x})]$ .

If  $\tau \geq \mu_i$  then this is equivalent to

$$\mu_i \hat{h}_\theta^+(\mathbf{x}^{i-}) - \tau \hat{h}_\theta^-(\mathbf{x}^{\rightarrow j}) + (\tau - \mu_i) \hat{h}_\theta^+(\mathbf{x}) < 0. \quad (22)$$

On the other hand, if  $\tau < \mu_i$  then (22) still applies except that  $(\tau - \mu_i)$  is negative and therefore  $\hat{h}_\theta^+(\mathbf{x})$  should be replaced with  $\hat{h}_\theta^-(\mathbf{x})$  in order to obtain an upper bound for the left-hand side. The confidence intervals are calculated using the method of reliability weights (Galassi et al. (2011), Sec. 21.7); please see Appendix F for full details. If there exists an action  $a^* \in A_{\mathbf{x}}$  which is better than all other actions according to the criteria described above, then this action is chosen by the improving policy. Otherwise, we cannot say with sufficient confidence which action should be chosen, and therefore as a ‘safety measure’ we set  $a^* = \theta(\mathbf{x})$ , so that the action prescribed by the base policy is chosen. By reverting to the base policy in the latter situation, we take a conservative approach and aim to ensure that improving actions are chosen only when they are statistically likely to be better than the action of the base policy, which helps to ensure that the policy given by approximate PI is a genuine improvement over the base policy. This is akin to the method of ‘safe policy improvement’, for which there is some existing literature (Laroche et al. (2019); Simão et al. (2023)).

After an action  $a^* \in A_{\mathbf{x}}$  has been chosen according to the rules above, the next transition of the system does not occur until a further  $\delta$  time units have elapsed, and we perform as many nested simulations as possible until the time limit  $\delta$  is reached. In our numerical experiments in Section 6 we set  $\delta = 0.01$  seconds, so that state transitions occur very frequently. Each nested simulation proceeds as follows: first we simulate a single transition of the system under the chosen action  $a^*$  and observe a new state  $\mathbf{x}'$ . Then, for each of the states  $\mathbf{y} \in F(\mathbf{x}')$ , we simulate the evolution of the system starting from  $\mathbf{y}$ , assuming that the base policy  $\theta$  is used for selecting actions, until eventually a state contained in the array  $U$  is reached. At this point an estimate is obtained for  $h_\theta(\mathbf{y})$  using equation (19), as in the offline part. The estimates  $\hat{h}_\theta(\mathbf{y})$  obtained from the nested simulations supplement the estimates already obtained from the offline part. Thus, the purpose of the nested simulations is to anticipate what the next state of the system might be (by sampling a new state  $\mathbf{x}'$ ) and enhance the PI policy’s ability to make a good choice of action at the next state (by improving the estimates  $\hat{h}_\theta(\mathbf{y})$  for states  $\mathbf{y} \in F(\mathbf{x}')$ ). The nested simulations are carried out repeatedly until  $\delta$  time units have elapsed, at which point the ‘actual’ realization of the next state  $\mathbf{x}'$  is observed and the main simulation continues.

By simulating a large number of time steps, with actions chosen using the rules above, we can evaluate the performance of the approximate PI policy. It should be noted that as the simulation progresses we continue to acquire new estimates  $\hat{h}_\theta(\mathbf{x})$  for states  $\mathbf{x}$  that are visited and the confidence intervals  $[\hat{h}_\theta^-(\mathbf{x}), \hat{h}_\theta^+(\mathbf{x})]$  become narrower, which means that we revert to the base policy less

often. Thus, the policy becomes stronger as we acquire more experience from interacting with the environment. This is a common theme in RL in general, as epitomized by the convergence properties of Q-learning (Watkins and Dayan (1992)) and the local improvement properties of other RL methods, such as policy gradient algorithms (Kakade (2001)).

## 6 Numerical results

In this section we report the results of computational experiments involving 1,000 randomly generated problem instances. Each problem instance includes a randomly generated network layout and also a randomly sampled set of system parameters. We generate the network layout using the method described in Tian and Shone (2026). Specifically, we construct a  $5 \times 5$  integer lattice with 25 nodes connected by horizontal and vertical edges. We then select a random subset of  $m$  nodes, where  $2 \leq m \leq 8$ , and designate these as machines. Thus, each machine  $i \in M$  is assigned coordinates  $(a_i, b_i)$  with  $a_i, b_i \in \{1, \dots, 5\}$ . The remaining (unselected) nodes serve as intermediate stages, although some of these may be redundant (in the sense that they would never be visited under an optimal policy) and can be ignored. Figure 3 shows an example in which machines are located at positions  $(1, 3)$ ,  $(2, 5)$  and  $(3, 1)$ . It is clear from the figure that, given these machine locations, the intermediate stages at  $(2, 1)$ ,  $(2, 2)$ ,  $(2, 3)$  and  $(2, 4)$  are the only ones that the repairer needs to visit in order to move between the machines in the most efficient manner possible, so there is no need to include intermediate stages at any other lattice points. The length of the shortest path between machines  $i$  and  $j$  is the Manhattan distance  $|a_i - a_j| + |b_i - b_j|$ .

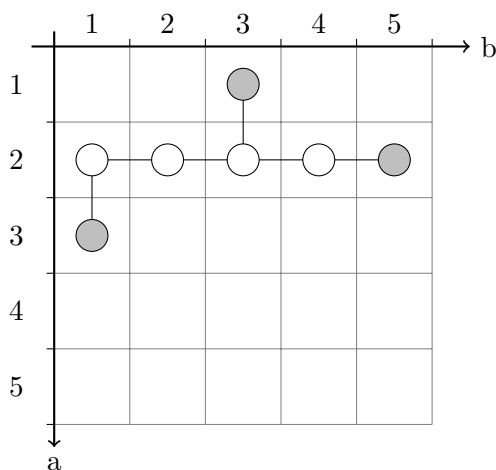


Figure 3: A network with machines at  $(1, 3)$ ,  $(2, 5)$  and  $(3, 1)$  and intermediate stages at  $(2, 1)$ ,  $(2, 2)$ ,  $(2, 3)$  and  $(2, 4)$ .

After the network layout has been generated for a specific problem instance, values of the system parameters are sampled randomly. We give a brief summary of this procedure here, but defer full details to Appendix G. We assume the number of possible degradation states,  $K_i$ , is the

same for each machine  $i \in M$ , but sample this value randomly from the set  $\{1, 2, 3, 4, 5\}$ . We also consider three different cases for cost functions  $f_i(x_i)$ , as follows: (i)  $f_i(x_i)$  increases linearly with the degradation state  $x_i$ ; (ii)  $f_i(x_i)$  is convex and increases quadratically with  $x_i$ ; (iii)  $f_i(x_i)$  increases linearly for  $x_i \in \{0, \dots, K_i - 1\}$ , but then increases much more sharply when  $x_i$  reaches the ‘failed’ state  $K_i$ . The type of cost function is the same for each machine within a particular problem instance, but the specific parameters used in the cost functions are machine-dependent, so some machines are associated with higher degradation costs than others. The degradation rates  $\lambda_i$  and repair rates  $\mu_i$  are randomly generated in such a way that the overall traffic intensity, given by  $\rho = \sum_i \lambda_i / \mu_i$ , lies within the range  $[0.1, 1.5]$ . Thus, we consider systems where the repairer is very under-utilized (i.e.  $\rho$  is near zero) and also systems where it is in very high demand (i.e.  $\rho > 1$ ). We also define the parameter  $\eta := \tau / (\sum_{i \in M} \lambda_i)$  as a measure of how fast the repairer is able to switch between adjacent nodes relative to the total of the machine degradation rates, and generate the value of  $\tau$  in order to consider a wide range of cases for the value of  $\eta$ . Please see Appendix G for full details of our parameter generation methods.

Recall that  $\theta^{\text{IND}}$  denotes the index policy defined in Section 4 (see Definition 4.4). We use  $\theta^{\text{OPI}}$  to denote the policy followed during implementation of the online policy improvement (OPI) method discussed in Section 5, with a time limit  $\delta = 0.01$  used for decision-making; that is, the system changes state every 0.01 seconds and the OPI heuristic is forced to make a decision within that time (please see Appendix F for further details of how the OPI heuristic is implemented in our experiments). Within each problem instance, we use simulation experiments to assess the performances of  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$ . Also, in instances with  $2 \leq m \leq 4$ , we use dynamic programming (specifically, exact policy improvement) to calculate optimal long-run average costs and rewards and use these to evaluate suboptimality percentages for these heuristics. As  $m$  increases, it becomes computationally infeasible to calculate an optimal policy using dynamic programming, so we do not calculate suboptimality percentages for systems with  $5 \leq m \leq 8$ .

Our experiments were carried out on an Intel(R) Core(TM) i5-6500T desktop computer, with 8GB of RAM, running Windows 10. The software used was Python 3.9.10, with the PyPy just-in-time compiler (<http://pypy.org>) used to speed up computations. During online implementation, the  $\theta^{\text{POL}}$  and  $\theta^{\text{IND}}$  policies require a negligible amount of time to make a decision at any individual time step. On the other hand,  $\theta^{\text{OPI}}$  carries out as many nested simulations as it can within the time limit  $\delta$  in order to improve the value function estimates for nearby states, as described in Section 5. Since all of our heuristics can make decisions quickly in an online setting, the number of experiments that we are able to perform is limited not by the decision-making processes of the heuristics themselves, but rather by the number of time steps that we need to simulate in order to accurately evaluate the performances of the heuristics. We simulate each heuristic over 500,000 time steps within each problem instance (please see Appendix I for further details).

The remainder of this section is organized as follows. In Section 6.1 we describe a simple heuristic based on polling systems that can be used to provide a useful benchmark for assessing the performances of the index and OPI heuristics. In Section 6.2, we evaluate the performances of

our heuristic policies by calculating their suboptimality relative to the optimal values in systems of modest size ( $2 \leq m \leq 4$ ) and also comparing them with the polling system policies. We also report the improvements that the OPI heuristic is able to achieve over the index heuristic in larger systems ( $5 \leq m \leq 8$ ). In Section 6.3 we investigate the effects of varying the system parameters on the performances of our heuristic policies.

### 6.1 Polling system heuristic

In order to provide an additional benchmark for comparisons, we introduce a simple ‘polling system’ heuristic that works by touring a subset of the  $m$  machines in a cyclic pattern and always fully repairing each machine it visits before moving to the next one. A polling system is a type of queueing model in which a single server provides service to multiple queues according to a pre-specified rule (Yechiali (1993); Levy and Sidi (2002); Boon et al. (2011)). Generally, it visits the queues cyclically in a fixed order and remains at each queue until a certain condition is met—which might be that the queue becomes empty (this is known as the ‘exhaustive’ regime) or that the queue no longer includes any jobs that were present when the server arrived (this is the ‘gated’ regime). Standard assumptions of polling systems include random job arrival and service times, and also a random switching time to move between any two queues. In our problem, we can develop a polling heuristic by interpreting machine degradation events as job arrivals and supposing that the repairer visits a subset of the machines in a fixed cyclic pattern and provides exhaustive service at each one. For each possible subset, the order is chosen by minimizing the total travel distance over one cycle. Specifically, consider a non-empty subset of machines  $\mathcal{M} \subseteq M$  and let  $\theta^{\text{POL}}(\mathcal{M})$  denote a policy under which the repairer visits the machines in a cyclic pattern  $(i_1, \dots, i_{|\mathcal{M}|}, i_1, \dots)$ , where  $i_1, \dots, i_{|\mathcal{M}|}$  are distinct nodes in  $\mathcal{M}$ . We assume that the order of visiting the machines is chosen in order to minimize the total travel distance,

$$D(i_1, \dots, i_{|\mathcal{M}|}) := \sum_{j=1}^{|\mathcal{M}|-1} d(i_j, i_{j+1}) + d(i_{|\mathcal{M}|}, i_1), \tag{23}$$

where  $d(i, j)$  is the length of a shortest path between machines  $i$  and  $j$ . In order to minimize this travel distance we must solve a ‘traveling salesman’ problem (TSP), but since we only implement the polling system heuristic for systems with  $m \leq 4$  the solution to the TSP is easy to find using brute force. Thus, for any given subset  $\mathcal{M} \subseteq M$  we calculate the performance of  $\theta^{\text{POL}}(\mathcal{M})$  by first solving a TSP to find a sequence  $(i_1, \dots, i_{|\mathcal{M}|})$  that minimizes the travel distance (23), then simulating the average cost incurred when the repairer follows the resulting cyclic pattern, enforcing the rule that damaged machines are always fully repaired before the repairer moves to the next machine. For any given system, we can obtain a useful benchmark for our index and OPI heuristics by simulating the performances of all possible ‘polling system’ policies  $\theta^{\text{POL}}(\mathcal{M})$  (taking into account all possible subsets  $\mathcal{M} \subseteq M$ ) and then reporting the lowest average cost among all of these. We use  $\theta^{\text{POL}} \in \arg \min_{\mathcal{M}} \{g_{\theta^{\text{POL}}(\mathcal{M})}(\mathbf{x})\}$  to denote the best-performing polling system heuristic within

a particular instance (with ties broken arbitrarily).

Although we use a polling heuristic primarily for benchmarking purposes, we note that heuristics based on polling systems have been successfully applied to dynamic routing problems based on order-picking in warehouses (Ran et al. (2020)) and mobile edge computing (Wang et al. (2023)). Various other applications of these methods are discussed in the survey by Borst and Boxma (2018).

## 6.2 Performances of heuristic policies in systems with $2 \leq m \leq 8$

In systems with  $2 \leq m \leq 4$  we measure the percentage suboptimalities of the heuristic policies  $\theta^{\text{POL}}$ ,  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  by comparing the average costs under these policies with the optimal values given by dynamic programming. In addition, we report the percentage suboptimalities when the average reward  $u_\theta(\mathbf{x})$  defined in (9) is used to measure the performance of a policy, rather than the average cost  $g_\theta(\mathbf{x})$ . Note that by Proposition 4.1 and Corollary 4.2 we have  $g_\theta(\mathbf{x}) = \sum_{j \in M} f_j(K_j) - u_\theta(\mathbf{x})$ , and therefore the *absolute* suboptimality is the same regardless of whether it is measured using costs or rewards. That is, for any policy  $\theta$  and  $\mathbf{x} \in S$  we have

$$g_\theta(\mathbf{x}) - g^* = u^* - u_\theta(\mathbf{x}),$$

where  $g^*$  and  $u^*$  denote the average cost and average reward, respectively, under the optimal policy. However, the *percentage* suboptimalities under the cost and reward formulations are  $100 \times (g_\theta(\mathbf{x}) - g^*)/g^*$  and  $100 \times (u^* - u_\theta(\mathbf{x}))/u^*$  respectively, and these are different because  $g^* \neq u^*$  in general. Although our original problem formulation in Section 3 was given in terms of costs, we would argue that measuring percentage suboptimalities in terms of rewards is also useful, because  $u^*$  is more robust than  $g^*$  with respect to the problem size (represented by the number of machines  $m$ ). To see this, note that as  $m$  increases, it becomes increasingly likely that some machines are never visited by the repairer; that is, they are left permanently broken. These non-visited machines cause the optimal average cost  $g^*$  to increase, but they have no effect on  $u^*$ . Hence, the ratio  $(g_\theta(\mathbf{x}) - g^*)/g^*$  decreases purely because of the inflation in  $g^*$ , whereas the ratio  $(u^* - u_\theta(\mathbf{x}))/u^*$  remains unaffected. In truth, both measures of suboptimality could be insightful depending on how one chooses to view the problem, so we consider both in this section.

To find the optimal values  $g^*$  and  $u^*$  in systems with  $2 \leq m \leq 4$ , the dynamic programming method we use is the policy iteration algorithm for unichain, average reward models, as described in Puterman (1994) (p. 378). This algorithm works by taking an initial base policy (which we set as the modified index policy  $\theta^{\text{M-IND}}$ , as discussed in Section 5) and then iterating between an ‘evaluation’ step and an ‘improvement’ step. The evaluation step is used to calculate the relative value function of the latest candidate policy, and the improvement step is used to find a superior policy by applying the Bellman equation at each state. The algorithm is guaranteed to converge to an optimal policy in a finite number of iterations. For details of the algorithm and associated computational times, please refer to Appendix H.

Tables 2 and 3 show, for each heuristic policy, 95% confidence intervals for the mean percentage

suboptimality relative to the optimal value  $g^*$  under the cost-based formulation and  $u^*$  under the reward-based formulation, respectively, based on results from 412 ‘small’ problem instances. We also report the 10th, 25th, 50th, 75th, and 90th percentiles of the percentage suboptimality distributions for each heuristic. These results indicate that the online policy improvement heuristic ( $\theta^{\text{OPI}}$ ) consistently outperforms the other two heuristics, with an average percentage suboptimality of 2.51% under the cost formulation and 1.02% under the reward formulation. The index heuristic ( $\theta^{\text{IND}}$ ) performs well in many instances, but is clearly inferior to  $\theta^{\text{OPI}}$ , most likely due to its short-sighted method of making decisions (as discussed in Section 4). It is unsurprising that  $\theta^{\text{OPI}}$  improves upon  $\theta^{\text{IND}}$  because the method of constructing  $\theta^{\text{OPI}}$  involves using  $\theta^{\text{IND}}$  as a base and then applying an improvement step. The polling system heuristic ( $\theta^{\text{POL}}$ ) is clearly weaker than  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$ , despite the fact that  $\theta^{\text{POL}}$  is a type of ‘ensemble’ heuristic that involves taking the best average cost (or reward) from all of the possible polling heuristics  $\theta^{\text{POL}}(\mathcal{M})$  within each problem instance. The weak performance of  $\theta^{\text{POL}}$  essentially shows the limitations of failing to respond dynamically to the latest machine degradation levels.

Table 2: Percentage suboptimalities of heuristic policies in 412 ‘small’ problem instances with  $2 \leq m \leq 4$ , evaluated under the cost-based formulation.

Heuristic	Mean	10th pct.	25th pct.	50th pct.	75th pct.	90th pct.
POL	$29.64 \pm 3.72$	0.00	5.78	17.36	35.25	71.14
IND	$8.11 \pm 1.04$	0.01	0.93	4.60	10.81	21.23
OPI	$2.51 \pm 0.55$	0.00	0.03	0.66	2.45	6.60

Table 3: Percentage suboptimalities of heuristic policies in 412 ‘small’ problem instances with  $2 \leq m \leq 4$ , evaluated under the reward-based formulation.

Heuristic	Mean	10th pct.	25th pct.	50th pct.	75th pct.	90th pct.
POL	$8.36 \pm 0.72$	0.00	2.07	7.43	12.64	18.95
IND	$3.01 \pm 0.45$	0.00	0.37	1.70	3.89	7.30
OPI	$1.02 \pm 0.32$	0.00	0.02	0.25	0.83	2.09

Tables 4 and 5 show the percentage improvements achieved by  $\theta^{\text{OPI}}$  against  $\theta^{\text{IND}}$  over all 1,000 problem instances, with the results categorized according to the number of machines  $m$ , under the cost and reward formulations respectively. In larger problem sizes (with  $m \geq 5$ ) it becomes impractical to compute exact percentage suboptimalities, but the results in these tables provide some assurance that the extra benefit given by applying online policy improvement remains fairly consistent as the size of the problem increases.

Table 4: Percentage improvements of OPI over the index policy for different values of  $m$ , evaluated under the cost-based formulation.

	$m = 2$ [132 instances]	$m = 3$ [148]	$m = 4$ [132]	$m = 5$ [146]	$m = 6$ [148]	$m = 7$ [149]	$m = 8$ [145]
OPI vs. IND	$3.45 \pm 1.13$	$5.43 \pm 1.13$	$5.00 \pm 0.91$	$5.96 \pm 0.97$	$5.12 \pm 0.76$	$6.33 \pm 1.02$	$4.98 \pm 0.94$

Table 5: Percentage improvements of OPI over the index policy for different values of  $m$ , evaluated under the reward-based formulation.

	$m = 2$ [132 instances]	$m = 3$ [148]	$m = 4$ [132]	$m = 5$ [146]	$m = 6$ [148]	$m = 7$ [149]	$m = 8$ [145]
OPI vs. IND	$1.83 \pm 1.23$	$2.97 \pm 1.53$	$2.12 \pm 0.42$	$2.37 \pm 0.37$	$2.32 \pm 0.42$	$2.12 \pm 0.36$	$2.08 \pm 0.39$

Table 6 shows, for each  $m \in \{2, \dots, 8\}$ , the percentage of time steps at which the heuristic  $\theta^{\text{OPI}}$  resorts to a ‘safe’ action choice by selecting the same action given by the index heuristic  $\theta^{\text{IND}}$ . Recall from Section 5 that, in the online implementation of the OPI heuristic, the policy makes a ‘safe’ choice of action if it cannot identify the best choice of action with a sufficient level of confidence. Table 6 shows that ‘safe’ action choices are selected with increasing frequency as  $m$  increases. This is due to the fact that, as the number of dimensions in the state space increases, it becomes more difficult to obtain reliable estimates for the values at individual states during the offline stage because there are relatively few states that are visited sufficiently often. This problem could be addressed by allocating more time to the offline stage as  $m$  becomes larger, or increasing the online time limit  $\delta$  (currently set at 0.01 seconds) so that the algorithm can gather more information about nearby states at each time step during the online stage.

Table 6: Percentage of ‘safe’ action choices under  $\theta^{\text{OPI}}$  for different values of  $m$ .

	$m = 2$ [132 instances]	$m = 3$ [148]	$m = 4$ [132]	$m = 5$ [146]	$m = 6$ [148]	$m = 7$ [149]	$m = 8$ [145]
Pct. of ‘safe’ action choices	$0.27 \pm 0.16$	$0.86 \pm 0.35$	$1.81 \pm 0.47$	$4.53 \pm 0.93$	$5.76 \pm 0.77$	$8.37 \pm 1.12$	$11.59 \pm 1.21$

### 6.3 Effects of varying the system parameters

Next, we investigate how the results from our experiments depend on the values of important system parameters. The parameters that we focus on are the traffic intensity  $\rho = \sum_{i \in M} \lambda_i / \mu_i$ , the relative switching speed  $\eta = \tau / (\sum_{i \in M} \lambda_i)$ , the type of cost function  $f_i(x_i)$  for  $i \in M$  and the maximum degradation state  $K$ . By varying these parameters we can investigate a wide range of different scenarios, including those in which degradation events occur very frequently (or infrequently) relative to the speed of repairs, those in which the repairer travels very slowly (or quickly) through the network, those in which costs increase superlinearly with the level of degradation, and

those in which there are several different stages of degradation before a machine fails completely. Thus, we aim to identify conditions under which our various heuristics perform well (or poorly). To the best of our knowledge, there are no publicly available benchmark instances that could be used to generate these parameter values, so we adopt a random generation method which ensures good coverage within the ranges of values that we consider; see Appendix G for full details.

In Tables 7 and 8 we categorize our results according to the traffic intensity,  $\rho = \sum_{i \in M} \lambda_i / \mu_i$ . The first three rows of the tables show percentage suboptimality of the heuristic policies  $\theta^{\text{POL}}$ ,  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  over the 412 instances with  $2 \leq m \leq 4$ , and the fourth rows show the percentage improvements of  $\theta^{\text{OPI}}$  over  $\theta^{\text{IND}}$  over all 1,000 instances with  $2 \leq m \leq 8$ . The columns correspond to disjoint, equal-length intervals for  $\rho$ . Within the table cells we report 95% confidence intervals for the relevant values under both the cost formulation (shown in regular font) and the reward formulation (shown in bold font). Table 9 follows the same format as Tables 7 and 8 except that the parameter of interest is  $\eta = \tau / (\sum_{i \in M} \lambda_i)$ , which measures the relative switching speed compared to the sum of the machine degradation rates. Table 10 again follows the same format except that we categorize the results according to the cost function  $f_i(x_i)$ . The three cost functions that we consider are: linear, quadratic and piecewise linear (see Appendix G for full details). Finally, Table 11 again follows the same format except that the parameter of interest is  $K$ , which is the maximum degradation state used for all machines in the system.

Table 7: Summary of numerical results categorized according to the value of  $\rho = \sum_i \lambda_i / \mu_i$  and  $0.1 \leq \rho < 0.7$ , with 95% confidence intervals shown under both the cost formulation (regular font) and the reward formulation (bold font)

	$0.1 \leq \rho < 0.3$	$0.3 \leq \rho < 0.5$	$0.5 \leq \rho < 0.7$
$\theta^{\text{POL}}$ Subopt. [412 instances]	72.03 ± 13.88 [ <b>9.11 ± 2.09</b> ]	42.98 ± 10.65 [ <b>9.91 ± 2.30</b> ]	24.16 ± 5.09 [ <b>8.66 ± 2.23</b> ]
$\theta^{\text{IND}}$ Subopt. [412 instances]	8.92 ± 2.31 [ <b>1.78 ± 0.73</b> ]	13.29 ± 3.87 [ <b>3.23 ± 1.24</b> ]	11.32 ± 3.70 [ <b>3.73 ± 1.37</b> ]
$\theta^{\text{OPI}}$ Subopt. [412 instances]	2.90 ± 1.27 [ <b>0.92 ± 0.79</b> ]	3.51 ± 1.63 [ <b>0.64 ± 0.31</b> ]	3.35 ± 1.54 [ <b>1.23 ± 0.91</b> ]
$\theta^{\text{OPI}}$ imp. vs. $\theta^{\text{IND}}$ [1,000 instances]	6.26 ± 1.02 [ <b>1.21 ± 0.55</b> ]	7.39 ± 1.34 [ <b>2.35 ± 0.71</b> ]	6.82 ± 1.04 [ <b>2.67 ± 0.50</b> ]

Table 8: Summary of numerical results categorized according to the value of  $\rho = \sum_i \lambda_i / \mu_i$  and  $0.7 \leq \rho < 1.5$ , with 95% confidence intervals shown under both the cost formulation (regular font) and the reward formulation (bold font)

	$0.7 \leq \rho < 0.9$	$0.9 \leq \rho < 1.1$	$1.1 \leq \rho < 1.3$	$1.3 \leq \rho < 1.5$
$\theta^{\text{POL}}$ Subopt. [412 instances]	19.57 $\pm$ 4.33 <b>[7.28 <math>\pm</math> 1.75]</b>	16.56 $\pm$ 4.10 <b>[8.96 <math>\pm</math> 1.97]</b>	13.96 $\pm$ 3.61 <b>[8.96 <math>\pm</math> 1.87]</b>	8.13 $\pm$ 1.66 <b>[6.27 <math>\pm</math> 1.21]</b>
$\theta^{\text{IND}}$ Subopt. [412 instances]	9.20 $\pm$ 2.91 <b>[2.95 <math>\pm</math> 1.43]</b>	6.46 $\pm$ 3.24 <b>[2.94 <math>\pm</math> 0.99]</b>	5.74 $\pm$ 1.80 <b>[4.30 <math>\pm</math> 1.81]</b>	2.86 $\pm$ 0.76 <b>[2.74 <math>\pm</math> 0.86]</b>
$\theta^{\text{OPI}}$ Subopt. [412 instances]	2.98 $\pm$ 1.51 <b>[0.80 <math>\pm</math> 0.36]</b>	2.87 $\pm$ 3.04 <b>[1.29 <math>\pm</math> 1.33]</b>	1.51 $\pm$ 0.63 <b>[1.04 <math>\pm</math> 0.51]</b>	0.87 $\pm$ 0.47 <b>[1.27 <math>\pm</math> 1.18]</b>
$\theta^{\text{OPI}}$ imp. vs. $\theta^{\text{IND}}$ [1,000 instances]	5.80 $\pm$ 1.05 <b>[2.49 <math>\pm</math> 1.09]</b>	3.81 $\pm$ 0.81 <b>[2.13 <math>\pm</math> 0.60]</b>	3.74 $\pm$ 0.75 <b>[3.06 <math>\pm</math> 1.29]</b>	2.61 $\pm$ 0.51 <b>[2.19 <math>\pm</math> 0.80]</b>

 Table 9: Summary of numerical results categorized according to the value of  $\eta = \tau / (\sum_{i \in M} \lambda_i)$ , with 95% confidence intervals shown under both the cost formulation (regular font) and the reward formulation (bold font)

	$0.1 \leq \eta < 0.4$	$0.4 \leq \eta < 0.7$	$0.7 \leq \eta < 1$	$1 \leq \eta < 4$	$4 \leq \eta < 7$	$7 \leq \eta < 10$
$\theta^{\text{POL}}$ Subopt. [412 instances]	5.89 $\pm$ 2.16 <b>[4.89 <math>\pm</math> 1.66]</b>	16.96 $\pm$ 5.09 <b>[10.50 <math>\pm</math> 2.38]</b>	23.78 $\pm$ 6.19 <b>[11.57 <math>\pm</math> 2.18]</b>	41.71 $\pm$ 11.45 <b>[10.72 <math>\pm</math> 1.61]</b>	37.87 $\pm$ 10.52 <b>[7.17 <math>\pm</math> 1.05]</b>	50.90 $\pm$ 10.82 <b>[5.88 <math>\pm</math> 0.82]</b>
$\theta^{\text{IND}}$ Subopt. [412 instances]	4.09 $\pm$ 1.82 <b>[4.26 <math>\pm</math> 1.92]</b>	5.33 $\pm$ 1.54 <b>[3.91 <math>\pm</math> 1.22]</b>	6.26 $\pm$ 2.27 <b>[3.62 <math>\pm</math> 0.97]</b>	10.43 $\pm$ 2.62 <b>[3.02 <math>\pm</math> 0.60]</b>	9.85 $\pm$ 3.15 <b>[2.03 <math>\pm</math> 0.53]</b>	12.43 $\pm$ 3.10 <b>[1.31 <math>\pm</math> 0.29]</b>
$\theta^{\text{OPI}}$ Subopt. [412 instances]	1.59 $\pm$ 1.07 <b>[2.18 <math>\pm</math> 1.55]</b>	1.24 $\pm$ 0.49 <b>[0.71 <math>\pm</math> 0.25]</b>	2.13 $\pm$ 1.11 <b>[1.35 <math>\pm</math> 0.87]</b>	2.84 $\pm$ 1.10 <b>[0.90 <math>\pm</math> 0.37]</b>	2.95 $\pm$ 1.45 <b>[0.55 <math>\pm</math> 0.22]</b>	4.20 $\pm$ 2.12 <b>[0.39 <math>\pm</math> 0.19]</b>
$\theta^{\text{OPI}}$ imp. vs. $\theta^{\text{IND}}$ [1,000 instances]	2.37 $\pm$ 0.78 <b>[3.26 <math>\pm</math> 1.66]</b>	4.09 $\pm$ 0.63 <b>[3.33 <math>\pm</math> 0.59]</b>	4.02 $\pm$ 0.89 <b>[2.58 <math>\pm</math> 0.60]</b>	7.76 $\pm$ 0.95 <b>[2.29 <math>\pm</math> 0.32]</b>	6.75 $\pm$ 1.15 <b>[1.27 <math>\pm</math> 0.22]</b>	5.83 $\pm$ 0.79 <b>[0.88 <math>\pm</math> 0.13]</b>

 Table 10: Summary of numerical results categorized according to the type of cost function  $f_i(x_i)$  used, with 95% confidence intervals shown under both the cost formulation (regular font) and the reward formulation (bold font)

	Linear	Quadratic	Piecewise linear
$\theta^{\text{POL}}$ Subopt. [412 instances]	26.47 $\pm$ 4.96 <b>[10.16 <math>\pm</math> 1.42]</b>	32.76 $\pm$ 7.41 <b>[7.55 <math>\pm</math> 1.07]</b>	29.09 $\pm$ 6.31 <b>[7.55 <math>\pm</math> 1.27]</b>
$\theta^{\text{IND}}$ Subopt. [412 instances]	5.38 $\pm$ 1.55 <b>[2.98 <math>\pm</math> 1.05]</b>	9.19 $\pm$ 1.99 <b>[2.82 <math>\pm</math> 0.59]</b>	9.50 $\pm$ 1.72 <b>[3.27 <math>\pm</math> 0.70]</b>
$\theta^{\text{OPI}}$ Subopt. [412 instances]	1.42 $\pm$ 0.54 <b>[0.95 <math>\pm</math> 0.49]</b>	3.20 $\pm$ 1.23 <b>[0.77 <math>\pm</math> 0.33]</b>	2.78 $\pm$ 0.83 <b>[1.38 <math>\pm</math> 0.80]</b>
$\theta^{\text{OPI}}$ imp. vs. $\theta^{\text{IND}}$ [1,000 instances]	3.58 $\pm$ 0.52 <b>[2.39 <math>\pm</math> 0.72]</b>	5.55 $\pm$ 0.64 <b>[2.19 <math>\pm</math> 0.30]</b>	6.49 $\pm$ 0.73 <b>[2.23 <math>\pm</math> 0.52]</b>

Table 11: Summary of numerical results categorized according to the maximum degradation state  $K$ , with 95% confidence intervals shown under both the cost formulation (regular font) and the reward formulation (bold font)

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$\theta^{\text{POL}}$ Subopt. [412 instances]	21.81 ± 5.95 [ <b>11.64 ± 2.29</b> ]	33.79 ± 11.74 [ <b>9.21 ± 1.81</b> ]	28.90 ± 7.34 [ <b>8.12 ± 1.51</b> ]	33.91 ± 8.89 [ <b>7.12 ± 1.30</b> ]	29.48 ± 7.74 [ <b>6.39 ± 1.04</b> ]
$\theta^{\text{IND}}$ Subopt. [412 instances]	3.26 ± 1.11 [ <b>2.86 ± 1.11</b> ]	4.74 ± 1.58 [ <b>2.01 ± 0.86</b> ]	6.94 ± 1.52 [ <b>2.85 ± 0.78</b> ]	12.01 ± 2.66 [ <b>3.66 ± 1.15</b> ]	12.06 ± 3.09 [ <b>3.42 ± 1.04</b> ]
$\theta^{\text{OPI}}$ Subopt. [412 instances]	1.19 ± 0.89 [ <b>1.57 ± 1.33</b> ]	0.96 ± 0.47 [ <b>0.50 ± 0.33</b> ]	1.83 ± 0.96 [ <b>0.95 ± 0.73</b> ]	4.10 ± 1.36 [ <b>1.07 ± 0.53</b> ]	3.93 ± 1.70 [ <b>0.98 ± 0.27</b> ]
$\theta^{\text{OPI}}$ imp. vs. $\theta^{\text{IND}}$ [1,000 instances]	2.16 ± 0.58 [ <b>1.71 ± 0.86</b> ]	4.33 ± 0.63 [ <b>1.89 ± 0.41</b> ]	5.08 ± 0.67 [ <b>2.18 ± 0.48</b> ]	7.02 ± 0.84 [ <b>2.75 ± 0.78</b> ]	7.03 ± 1.09 [ <b>2.70 ± 0.73</b> ]

From Tables 7 and 8 we can see that the polling system heuristic  $\theta^{\text{POL}}$  is very weak when  $\rho$  is small, but improves as  $\rho$  increases. Indeed, when  $\rho$  is small, we often have a situation where only one machine is degraded (that is,  $x_j > 0$  for only one  $j \in M$ ) and the repairer should go directly to that machine under an optimal policy, but the polling heuristic forces the repairer to make wasteful visits to all machines according to a fixed cyclic pattern. The other heuristics  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  also seem to achieve lower suboptimality values as  $\rho$  increases under the cost formulation, but this may be caused by the inflation in  $g^*$  that occurs when the overall degradation burden increases (as discussed earlier). The bold-font figures in Tables 7 and 8 show that when the reward formulation is used, there is no clear trend for the performances of  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  to get better or worse as  $\rho$  increases. Instead,  $\theta^{\text{IND}}$  is consistently within 5% of optimality, while  $\theta^{\text{OPI}}$  is consistently within 2%.

Table 9 shows that, for all three heuristic policies, there is a general trend for suboptimality to increase with  $\eta$  when the cost formulation is used. The same reasoning given above also applies in this case, since smaller  $\eta$  values imply that machine degradation happens more frequently (relative to the repairer’s movement speed), so the optimal value  $g^*$  will tend to increase. Under the reward formulation, however, it seems that the trend for  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  is the opposite: as  $\eta$  increases, these heuristics seem to improve their performance. This may be due to the fact that the indices used by  $\theta^{\text{IND}}$  (which also influence the decisions of  $\theta^{\text{OPI}}$ ) are calculated under the assumption that the repairer will move directly to a particular demand point without changing direction, which is more likely to be an optimal course of action if  $\eta$  is large (since there is less chance of new degradation events occurring while the repairer is moving).

Table 10 shows that under the cost formulation, the three heuristic policies seem to achieve the lowest suboptimality values when the linear cost function is used. However, under the reward formulation, this trend is no longer apparent. Recall that the linear and quadratic cost functions are defined by  $f_i(x_i) = c_i x_i$  and  $f_i(x_i) = c_i x_i^2$  respectively, whereas the piecewise linear cost function  $f_i(x_i) = c_i(x_i + 10 \cdot \mathbb{I}(x_i = K_i))$  assumes an extra penalty for any machine that has reached the

state of complete failure. One might imagine that the piecewise linear case would be a difficult case to optimize, as even the machines with relatively low service priorities should still (ideally) be visited regularly enough to make complete failure unlikely. From Table 10, however, there is no clear evidence that the heuristics perform worse in this case than in the other two.

Finally, Table 11 shows that the effect of measuring machine degradation on a finer scale (in other words, increasing the number of possible degradation states) is rather difficult to pinpoint based on the suboptimality percentages. When the cost formulation is used, the performances of  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  appear to become slightly worse as  $K$  increases, which may be expected since the problem becomes intrinsically more complex. However, under the reward formulation, this trend seems to disappear. Indeed, even in the  $K = 5$  case,  $\theta^{\text{IND}}$  and  $\theta^{\text{OPI}}$  are able to achieve mean suboptimalities of less than 3.5% and 1% respectively. Encouragingly, in all of the cases shown in Tables 7-11,  $\theta^{\text{OPI}}$  is able to obtain healthy improvements over  $\theta^{\text{IND}}$ .

## 7 Conclusions

This paper considers an original model of a stochastic, dynamic repair and maintenance problem in which a single repairer provides service to machines that are spatially dispersed on a network. By allowing switching and service times to be interruptible, we obtain an MDP in which decisions can be made at any point in time. Each machine evolves through a finite set of degradation states, and switching times depend not only on the machines involved but also on the progress made during any interrupted operations. Costs are linked to machine conditions using functions that can be defined arbitrarily. Although we have presented the model in the context of machine maintenance, our mathematical model is quite general and could be applied to other settings such as cybersecurity and healthcare routing, as discussed in Section 2.

From the methodological point of view, our main contributions are (i) the development of an index heuristic which exploits the interruptibility of service and switching operations by recalculating indices at every time step, and (ii) the use of a rollout-based RL method, with a novel safety feature based on reliability weights, to improve upon the index heuristic by making decisions in a less myopic way. The index heuristic is simple to implement and can be scaled to large problem instances without difficulty, as the number of indices to be calculated at each time step increases only linearly with the number of machines. Furthermore, our results in Section 4 have proven that this heuristic is optimal in certain special cases. Our numerical results in Section 6 have shown that the rollout-based online policy improvement method is close to optimal in small instances and consistently yields improvements over the index heuristic in larger instances, despite being restricted to only 0.01 seconds of simulation time per system transition during the online phase. This suggests that the approach is suitable for use in fast-changing real-world systems, such as those encountered in communications and computing.

Future research could focus on adapting these methods to more general frameworks, such as

those with multiple repairers or nonstationary dynamics. It would be interesting to explore settings with partial observability, in which the repairer has incomplete knowledge of the conditions of the various machines. This could lead to the formulation of a partially observable MDP or POMDP (Kaelbling et al. (1998)). Another interesting avenue to explore, which could be relevant to security settings, is where degradation events occur as a result of adversarial behavior by a strategic opponent, rather than occurring at random. This would lead to a game theoretical formulation. It would also be interesting to investigate the use of other RL approaches, such as policy function approximation, which might be faster to implement in the online stage.

## Acknowledgments.

The first author was funded by a PhD studentship from the Engineering and Physical Sciences Research Council (EPSRC), under grant EP/W523811/1.

## References

- Abbou, A. and Makis, V. (2019). Group maintenance: A restless bandits approach. *INFORMS Journal on Computing*, 31(4):719–731.
- Ahmad, W., Hasan, O., Pervez, U., and Qadir, J. (2017). Reliability modeling and analysis of communication networks. *Journal of Network and Computer Applications*, 78:191–215.
- Argon, N., Ding, L., Glazebrook, K., and Ziya, S. (2009). Dynamic routing of customers with general delay costs in a multiserver queuing system. *Probability in the Engineering and Informational Sciences*, 23(2):175–203.
- Arora, A., Krishnan, R., Telang, R., and Yang, Y. (2010). An empirical analysis of software vendors’ patch release behavior: impact of vulnerability disclosure. *Information Systems Research*, 21(1):115–132.
- Ayesta, U., Gupta, M., and Verloop, I. (2021). On the computation of Whittle’s index for Markovian restless bandits. *Mathematical Methods of Operations Research*, 93:179–208.
- Bellman, R. (1954). The Theory of Dynamic Programming. *Bulletin of the American Mathematical Society*, 60:503–516.
- Berg, M. (1976). Optimal replacement policies for two-unit machines with increasing running costs 1. *Stochastic Processes and their Applications*, 4(1):89–106.
- Bertsekas, D. (2020). *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. Athena Scientific.
- Boon, M., van der Mei, R., and Winands, E. (2011). Applications of polling systems. *Surveys in Operations Research and Management Science*, 16:35–57.
- Borst, S. and Boxma, O. (2018). Polling: past, present, and perspective. *TOP*, 26(3):335–369.
- Cancela, H., Murray, L., Robledo, F., Romero, P., and Sartor, P. (2022). On the reliability estimation of stochastic binary systems. *International Transactions in Operational Research*, 29(3):1688–1722.

- Cho, D. and Parlar, M. (1991). A survey of maintenance models for multi-unit systems. *European Journal of Operational Research*, 51(1):1–23.
- Cinlar, E. (1975). *Introduction to Stochastic Processes*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Cire, A. A. and Diamant, A. (2022). Dynamic scheduling of home care patients to medical providers. *Production and Operations Management*, 31(11):4038–4056.
- Dekker, R. and Roelvink, I. (1995). Marginal cost criteria for preventive replacement of a group of components. *European Journal of Operational Research*, 84(2):467–480.
- Dekker, R., Wildeman, R., and Van der Duyn Schouten, F. (1997). A review of multi-component maintenance models with economic dependence. *Mathematical Methods of Operations Research*, 45(3):411–435.
- Demirci, E., Arts, J., and Houtum, G. (2025). A restless bandit approach for capacitated condition based maintenance scheduling. *Flexible Services and Manufacturing Journal*, 37(1):179–207.
- Dey, D., Lahiri, A., and Zhang, G. (2015). Optimal policies for security patch management. *INFORMS Journal on Computing*, 27(3):462–477.
- Duenyas, I. and Van Oyen, M. (1996). Heuristic Scheduling of Parallel Heterogeneous Queues with Set-Ups. *Management Science*, 42(6):814–829.
- Fairley, L., Shone, R., Jacko, P., and Huang, J. (2025). A bi-objective markov decision process design approach to redundancy allocation with dynamic maintenance for a parallel system. *European Journal of Operational Research (in press)*.
- Fielder, A., Panaousis, E., Malacaria, P., Hankin, C., and Smeraldi, F. (2016). Decision support approaches for cyber security investment. *Decision Support Systems*, 86:13–23.
- Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Rossi, F. (2011). GNU Scientific Library Reference Manual-Edition 1.15, for GSL Version 1.15.
- Geurtsen, M., Didden, J. B., Adan, J., Atan, Z., and Adan, I. (2023). Production, maintenance and resource scheduling: A review. *European Journal of Operational Research*, 305(2):501–529.
- Glazebrook, K., Mitchell, H., and Ansell, P. (2005). Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research*, 165(1):267–284.
- Hernandez-Leal, P., Kartal, B., and Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797.
- Horák, K., Bošanský, B., Kovařík, V., and Kiekintveld, C. (2023). Solving zero-sum one-sided partially observable stochastic games. *Artificial Intelligence*, 316:103838.
- Ioannidis, C., Pym, D., and Williams, J. (2012). Information security trade-offs and optimal patching policies. *European Journal of Operational Research*, 216(2):434–444.
- Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134.
- Kakade, S. M. (2001). A natural policy gradient. *Advances in Neural Information Processing Systems*, 14.

- Khorasanian, D., Patrick, J., and Sauré, A. (2024). Dynamic home care routing and scheduling with uncertain number of visits per referral. *Transportation Science*, 58(4):841–859.
- Krishnan, K. (1990). Joining the right queue: A state-dependent decision rule. *IEEE Transactions on Automatic Control*, 35:104–108.
- Lagos, G. and Romero, P. (2020). On the reliability of dynamical stochastic binary systems. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 516–527.
- Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.
- Lee, C.-Y. and Lin, C.-S. (2001). Single-machine scheduling with maintenance and repair rate-modifying activities. *European Journal of Operational Research*, 135(3):493–513.
- Lee, S. and Lee, Y. H. (2020). Improving emergency department efficiency by patient scheduling using deep reinforcement learning. In *Healthcare*, volume 8, page 77.
- Levy, H. and Sidi, M. (2002). Polling systems: applications, modeling, and optimization. *IEEE Transactions on Communications*, 38(10):1750–1760.
- Lippman, S. (1975). Applying a new device in the optimisation of exponential queueing systems. *Operations Research*, 23(4):687–710.
- Liu, X., Liu, Y., Sun, Z., Wang, B., and Zhao, Y. (2025). Dynamic optimization of maintenance strategy for bridges in regional transportation network through semi-markov processes. *Automation in Construction*, 174:106175.
- Luo, F., Xu, T., Li, J., and Xu, F. (2025). MDP-AD: A Markov decision process-based adaptive framework for real-time detection of evolving and unknown network attacks. *Alexandria Engineering Journal*, 126:480–490.
- Medina, G. M., Castillo-Villar, K. K., and Bhuiyan, T. H. (2025). Integrating IT and OT for Cybersecurity: A Stochastic Optimization Approach via Attack Graphs. *IEEE Access*.
- Mookerjee, R. and Samuel, J. (2023). Managing the security of information systems with partially observable vulnerability. *Production and Operations Management*, 32(9):2902–2920.
- Mosheiov, G. and Oron, D. (2021). A note on scheduling a rate modifying activity to minimize total late work. *Computers & Industrial Engineering*, 154:107138.
- Naderi, B., Begen, M. A., Zaric, G. S., and Roshanaei, V. (2023). A novel and efficient exact technique for integrated staffing, assignment, routing, and scheduling of home care services under uncertainty. *Omega*, 116:102805.
- Nasir, J. A. and Kuo, Y.-H. (2024). Stochastic home care transportation with dynamically prioritized patients: An integrated facility location, fleet sizing, and routing approach. *Transportation Research Part B: Methodological*, 184:102949.
- Nicolai, R. and Dekker, R. (2008). Optimal maintenance of multi-component systems: a review. In Kobbacy, K. and Murthy, D., editors, *Complex System Maintenance Handbook*, pages 263–286. Springer Science & Business Media.

- Nikzad, E., Bashiri, M., and Abbasi, B. (2021). A matheuristic algorithm for stochastic home health care planning. *European Journal of Operational Research*, 288(3):753–774.
- Nino-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98.
- Ogunfowora, O. and Najjaran, H. (2023). Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization. *Journal of Manufacturing Systems*, 70:244–263.
- Ohashi, M. (1982). Group replacement policy for a maintained coherent system. *Journal of the Operations Research Society of Japan*, 25(3):228–237.
- Özekici, S. (1988). Optimal periodic replacement of multicomponent reliability systems. *Operations Research*, 36(4):542–552.
- Patrick, J., Puterman, M. L., and Queyranne, M. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525.
- Poonia, P. K. (2021). Performance assessment of a multi-state computer network system in series configuration using copula repair. *International Journal of Reliability and Safety*, 15(1-2):68–88.
- Powell, W. (2022). *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*. Wiley.
- Puterman, M. (1994). *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. Wiley & Sons, New York.
- Ran, W., Liu, S., and Zhang, Z. (2020). A polling-based dynamic order-picking system considering priority orders. *Complexity*, 2020(1):4595316.
- Ruiz-Hernandez, D., Pinar-Perez, J., and Delgado-Gomez, D. (2020). Multi-machine preventive maintenance scheduling with imperfect interventions: A restless bandit approach. *Computers & Operations Research*, 119:104927.
- Serfozo, R. (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research*, 27(3):616–620.
- Seyedshohadaie, S. R., Damnjanovic, I., and Butenko, S. (2010). Risk-based maintenance and rehabilitation decisions for transportation infrastructure networks. *Transportation Research Part A: Policy and Practice*, 44(4):236–248.
- Shone, R., Knight, V., and Harper, P. (2020). A conservative index heuristic for routing problems with multiple heterogeneous service facilities. *Mathematical Methods of Operations Research*, 92:511–543.
- Simão, T. D., Suilen, M., and Jansen, N. (2023). Safe policy improvement for pomdps via finite-state controllers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15109–15117.
- Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction (2nd edition)*. The MIT Press.

- Tao, W., Lin, P., and Wang, N. (2021). Optimum life-cycle maintenance strategies of deteriorating highway bridges subject to seismic hazard by a hybrid markov decision process model. *Structural Safety*, 89:102042.
- Tian, D. and Shone, R. (2026). Stochastic dynamic job scheduling with interruptible setup and processing times: An approach based on queueing control. *European Journal of Operational Research*, 329(3):920–934.
- Tijms, H. (2003). *A First Course in Stochastic Models*. Wiley & Sons, Chichester.
- Wang, X., Yang, Z., and Ding, H. (2023). Application of polling scheduling in mobile edge computing. *Axioms*, 12(7):709.
- Ward, T., Jenab, K., Ortega-Moody, J., and Staub, S. (2024). A comprehensive review of machine learning techniques for condition-based maintenance. *International Journal of Prognostics and Health Management*, 15(2).
- Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3):279–292.
- Whittle, P. (1988). Restless Bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25:287–298.
- Wijnmalen, D. and Hontelez, J. (1997). Coordinated condition-based repair strategies for components of a multi-component maintenance system with discounts. *European Journal of Operational Research*, 98(1):52–63.
- Woo, Y.-B., Jung, S., and Kim, B. S. (2017). A rule-based genetic algorithm with an improvement heuristic for unrelated parallel machine scheduling problem with time-dependent deterioration and multiple rate-modifying activities. *Computers & Industrial Engineering*, 109:179–190.
- Xin, G., Liang, Z., Hu, Y., Long, G., Zhang, Y., and Liang, P. (2023). Modeling the optimal maintenance strategy for bridge elements based on agent sequential decision making. *Applied Sciences*, 14(1):14.
- Xu, G. and Azhari, F. (2022). Data-driven optimization of repair schemes and inspection intervals for highway bridges. *Reliability Engineering & System Safety*, 228:108779.
- Xu, G. and Guo, F. (2022). Sustainability-oriented maintenance management of highway bridge networks based on Q-learning. *Sustainable Cities and Society*, 81:103855.
- Yechiali, U. (1993). *Analysis and control of polling systems*. In L. Donatiello and R. Nelson (Eds), Performance Evaluation of Computer and Communication Systems, Performance 1993, Sigmetrics 1993. Lecture Notes in Computer Science, vol 729. Springer, Berlin, Heidelberg.
- Yousefi, N., Tsianikas, S., and Coit, D. W. (2020). Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components. *Quality Engineering*, 32(3):388–408.
- Żebrowski, P., Couce-Vieira, A., and Mancuso, A. (2022). A bayesian framework for the analysis and optimal mitigation of cyber threats to cyber-physical systems. *Risk Analysis*, 42(10):2275–2290.
- Zhang, K., Yang, Z., and Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384.

## Online Appendices

### A Proof of Proposition 3.1.

We prove the result by showing that our finite-state MDP belongs to the *communicating* class of MDP models, as defined in Puterman (1994) (p. 348). To this end, it will be sufficient to prove the lemma below.

**Lemma A.1.** *For any two states  $\mathbf{x}, \mathbf{x}' \in S$  there exists a stationary policy  $\theta$  such that  $\mathbf{x}'$  is accessible from  $\mathbf{x}$  in the Markov chain induced by  $\theta$ .*

The lemma is proved as follows. Let  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and  $\mathbf{x}' = (i', (x'_1, \dots, x'_m))$  be two given states in  $S$ . Let  $\theta$  be a stationary policy that chooses an action under state  $\mathbf{x}$  as follows:

- If  $x_1 = x_2 = \dots = x_m = 0$  then we identify the shortest path from  $i$  to  $i'$  and attempt to take the first step along that path. (If  $i = i'$  then we just remain at  $i$ .)
- If  $x_j > 0$  for at least one machine  $j \in M$  then we identify the ‘most-damaged’ machine, i.e. the machine  $j$  with  $x_j \geq x_l$  for all  $l \in M$ . In case of ties, we assume that  $j$  is selected according to some pre-determined priority ordering of the machines. If the repairer is already at node  $j$  (i.e.  $j = i$ ), then it remains there. Otherwise, the repairer identifies the shortest path from its current location  $i$  to node  $j$ , and attempts to take the first step along that path.

For clarity, the ‘shortest path’ between two nodes  $i, i' \in V$  means the shortest possible sequence of connected nodes from  $i$  to  $i'$ . In case two or more paths are tied for the shortest distance, the path can be chosen according to some fixed priority ordering of the nodes. Also, note that the rules above are implemented dynamically at each time step, so it may happen (for example) that the repairer begins moving towards a particular machine  $j$  with  $x_j \geq x_l$  for all  $l \in M$  but, during the switching process, another machine  $r$  degrades to the point that  $x_j < x_r$  and therefore the repairer changes course and moves towards node  $r$  instead.

To prove the proposition, it suffices to show that there is a possible sequence of random events that causes the system to reach state  $\mathbf{x}'$  when the policy  $\theta$  described above is followed. In the first case ( $x_1 = x_2 = \dots = x_m = 0$ ), it is clear that the state  $(i', (0, 0, \dots, 0))$  can be reached if none of the machines degrade before the repairer arrives at node  $i'$ . If the system is in state  $(i', (0, 0, \dots, 0))$  and one of the machines (say machine 1) degrades, then the system transitions to state  $(i', (1, 0, \dots, 0))$  and according to the rules above, the repairer should attempt to move to the first node on the shortest path from  $i'$  to 1. However, with probability  $1 - \tau$ , the switch ‘fails’ and the repairer remains at  $i'$ . Furthermore, it is possible for other machines to deteriorate (possibly multiple times) before the switch eventually succeeds. (It may also happen that  $i' = 1$ , in which case the repairer remains at machine 1 and attempts to repair it, but degradations happen while the repair

is in progress.) Using this reasoning, we can see that the state  $\mathbf{x}' = (i', x'_1, \dots, x'_m)$  (where each  $x'_j$  is an arbitrary value in  $\{0, \dots, K_j\}$ ) is accessible from  $(i, (0, 0, \dots, 0))$  via an appropriate sequence of random degradation events.

In the second case ( $x_j > 0$  for at least one machine  $j \in M$ ), suppose that no further degradations occur until the repairer has restored all of the machines to their ‘pristine’ states. This fortuitous sequence of events allows the system to eventually transition to state  $(j, (0, 0, \dots, 0))$ , where in this case  $j \in M$  denotes the last machine to be fully repaired. We can then simply repeat the arguments given for the previous case, replacing  $i$  with  $j$ , in order to show that state  $\mathbf{x}' = (i', (x'_1, \dots, x'_m))$  can eventually be reached. This completes the proof of the lemma, and the result follows from the theory in Puterman (1994) (see Theorem 8.3.2, p. 351).  $\square$

## B Proof of Proposition 4.1.

In this proof we consider an arbitrary stationary policy,  $\theta$ . Since  $\theta$  is not necessarily optimal, it may be the case that the average cost  $g_\theta(\mathbf{x})$  depends on the initial state  $\mathbf{x}$ . Let  $P_\theta$  denote the transition matrix under  $\theta$ , consisting of the one-step transition probabilities  $p_{\mathbf{x},\mathbf{y}}(\theta(\mathbf{x}))$  defined in (1), where  $\theta(\mathbf{x})$  is the action chosen at state  $\mathbf{x}$  under  $\theta$ . Since the state space  $S$  is finite and  $P_\theta$  is a stochastic matrix, it follows from Proposition 8.1.1 in Puterman (1994) (p. 333) that

$$g_\theta(\mathbf{x}) = \sum_{\mathbf{y} \in S} \pi_\theta(\mathbf{x}, \mathbf{y}) c(\mathbf{y}),$$

where  $\pi_\theta(\mathbf{x}, \mathbf{y}) = \lim_{k \rightarrow \infty} \mathbb{P}_\theta(\mathbf{x}_k = \mathbf{y} \mid \mathbf{x}_0 = \mathbf{x})$  denotes the limiting probability of moving from state  $\mathbf{x}$  to state  $\mathbf{y}$  in  $k$  transitions under policy  $\theta$ , and this probability can be obtained as the entry in the row corresponding to state  $\mathbf{x}$  and the column corresponding to state  $\mathbf{y}$  in the matrix  $P_\theta^* := \lim_{k \rightarrow \infty} (P_\theta)^k$ . It follows that equation (7) is equivalent to

$$\sum_{\mathbf{y} \in S} \pi_\theta(\mathbf{x}, \mathbf{y}) c(\mathbf{y}) = \sum_{\mathbf{y} \in S} \pi_\theta(\mathbf{x}, \mathbf{y}) \tilde{c}(\mathbf{y}, \theta(\mathbf{y})) \quad \forall \mathbf{x} \in S. \tag{24}$$

Let the initial state  $\mathbf{x}$  be fixed. For each machine  $j \in M$  and  $k \in \{1, 2, \dots, K_j\}$ , let  $G_{j,k}$  denote the set of states  $\mathbf{y}$  under which the repairer is at node  $j$ ,  $y_j = k$  and the action chosen under  $\theta$  is to remain at node  $j$ . By considering the costs incurred at each of the individual machines, it may be seen that equation (24) can be expressed as

$$\sum_{j \in M} \sum_{\mathbf{y} \in S} \pi_\theta(\mathbf{x}, \mathbf{y}) f_j(y_j) = \sum_{j \in M} f_j(K_j) - \sum_{j \in M} \sum_{k=1}^{K_j} \sum_{\mathbf{y} \in G_{j,k}} \pi_\theta(\mathbf{x}, \mathbf{y}) \frac{\mu_j}{\lambda_j} [f_j(K_j) - f_j(k-1)].$$

It will therefore be sufficient to show that, for an arbitrary fixed  $j \in M$ , we have

$$\sum_{\mathbf{y} \in S} \pi_{\theta}(\mathbf{x}, \mathbf{y}) f_j(y_j) = f_j(K_j) - \sum_{k=1}^{K_j} \sum_{\mathbf{y} \in G_{j,k}} \pi_{\theta}(\mathbf{x}, \mathbf{y}) \frac{\mu_j}{\lambda_j} [f_j(K_j) - f_j(k-1)]. \quad (25)$$

Using the detailed balance equations for ergodic Markov chains (see, for example, Cinlar (1975)), the rate at which machine  $j$  transitions from state  $y_j$  to  $y_{j+1}$  under  $\theta$  must equal the rate at which it transitions from  $y_{j+1}$  to  $y_j$ . This implies that

$$\lambda_j \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) = \mu_j \sum_{\mathbf{y} \in G_{j,k+1}} \pi_{\theta}(\mathbf{x}, \mathbf{y}) \quad \forall k \in \{0, 1, \dots, K_j - 1\}.$$

Therefore the right-hand side of (25) can be written equivalently as

$$f_j(K_j) - \sum_{k=0}^{K_j-1} \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) [f_j(K_j) - f_j(k)]. \quad (26)$$

Then, using the fact that

$$\sum_{k=0}^{K_j} \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) = 1,$$

we can write (26) in the form

$$\begin{aligned} & \sum_{k=0}^{K_j} \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) f_j(K_j) - \sum_{k=0}^{K_j-1} \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) [f_j(K_j) - f_j(k)] \\ &= \sum_{\mathbf{y} \in S: y_j = K_j} \pi_{\theta}(\mathbf{x}, \mathbf{y}) f_j(K_j) + \sum_{k=0}^{K_j-1} \sum_{\mathbf{y} \in S: y_j = k} \pi_{\theta}(\mathbf{x}, \mathbf{y}) f_j(k) \\ &= \sum_{\mathbf{y} \in S} \pi_{\theta}(\mathbf{x}, \mathbf{y}) f_j(y_j) \end{aligned}$$

which equals the left-hand side of (25) as required. This confirms that the average costs for machine  $j$  are equal under both cost formulations, and since this argument can be repeated for each machine  $j \in M$  and initial state  $\mathbf{x} \in S$ , the proof is complete.  $\square$

## C Proof of Proposition 4.5

We assume that (i) there are  $m$  machines arranged in a complete graph with no intermediate stages, i.e.,  $V = M$ ; (ii) for all  $j \in M$ ,  $K_j = 1$ ; (iii) all  $m$  machines have identical parameter settings.

Under these conditions, the state space has the form

$$S = \{(i, (x_1, \dots, x_m)) \mid i \in M, x_j \in \{0, 1\} \text{ for } j \in M\}.$$

Machine  $i \in M$  is working if  $x_i = 0$ , and has failed if  $x_i = 1$ . Throughout this proof we omit the subscripts on the degradation rates  $\lambda_j$ , the repair rates  $\mu_j$  and the cost functions  $f_j$ , since the third assumption implies that these are the same for all  $j \in M$ . Also, given that each machine has only two possible conditions, we will depart from our usual notational convention and use  $k$  as a time step index rather than an indicator of a machine's condition.

The general strategy of this proof is to determine the decisions made by the index policy  $\theta^{\text{IND}}$  under these conditions, and then show that these coincide with the decisions made by an optimal policy. First, suppose that the system is operating under the index policy. We can consider three possible cases: (a) the repairer is at a node  $i$  with  $x_i = 1$ ; (b) the repairer is at a node  $i$  with  $x_i = 0$  and there is at least one node  $j \neq i$  with  $x_j = 1$ ; (c) the repairer is at a node  $i$  with  $x_i = 0$  and we also have  $x_j = 0$  for all  $j \neq i$ . In case (a), the repairer has a choice between remaining at the failed machine  $i$  or switching to an alternative machine. Recall that the index policy makes decisions using the indices  $\Phi_i^{\text{stay}}(x_i)$ ,  $\Phi_{ij}^{\text{move}}(x_j)$  and  $\Phi_{ij}^{\text{wait}}(x_j)$  defined in (12), (13) and (14) respectively. Given the assumption that  $K_j = 1$  for all  $j \in S$ , the index for remaining at node  $i$  simplifies to

$$\Phi_i^{\text{stay}}(x_i) = \Phi_i^{\text{stay}}(1) = \frac{\mathbb{E}[R_i(1)]}{\mathbb{E}[T_i(1)]} = \frac{\mu}{\lambda} f(1), \quad (27)$$

which is simply equal to the reward rate earned while the machine is being repaired. Next, consider the index for moving to an alternative node  $j \neq i$ . According to the rules of the heuristic, the repairer should only consider switching to nodes belonging to the set  $J$  defined in (16). Thus, we need to check whether  $\Phi_{ij}^{\text{move}}(x_j) \geq \Phi_{ij}^{\text{wait}}(x_j)$  for each  $j \neq i$ . We can express these indices as follows:

$$\Phi_{ij}^{\text{move}}(x_j) = \begin{cases} \mathbb{P}(X_j = 1) \frac{\mathbb{E}[R_j(1)]}{\mathbb{E}[D_{ij}|X_j = 1] + \mathbb{E}[T_j(1)]} = \frac{\mu\tau f(1)}{\tau^2 + (2\mu + \lambda)\tau + \mu\lambda}, & \text{if } x_j = 0, \\ \frac{\mathbb{E}[R_j(1)]}{\mathbb{E}[D_{ij}] + \mathbb{E}[T_j(1)]} = \left(\frac{\tau}{\mu + \tau}\right) \frac{\mu}{\lambda} f(1), & \text{if } x_j = 1, \end{cases} \quad (28)$$

$$\Phi_{ij}^{\text{wait}}(x_j) = \frac{\mathbb{E}[R_j(1)]}{1/\lambda + \mathbb{E}[D_{ij}] + \mathbb{E}[T_j(1)]} = \frac{\mu\tau f(1)}{(\mu + \lambda)\tau + \mu\lambda}. \quad (29)$$

By comparing the expressions in (28) and (29) we can see that  $\Phi_{ij}^{\text{move}}(x_j) < \Phi_{ij}^{\text{wait}}(x_j)$  if  $x_j = 0$ , but  $\Phi_{ij}^{\text{move}}(x_j) > \Phi_{ij}^{\text{wait}}(x_j)$  if  $x_j = 1$ . Therefore the repairer only considers switching to node  $j$  if  $x_j = 1$ . However, even in this case  $\Phi_{ij}^{\text{move}}(x_j)$  is smaller than the index  $\Phi_i^{\text{stay}}(x_i)$  in (27) (noting that  $R_i(1) \stackrel{\text{dist}}{=} R_j(1)$  and  $T_i(1) \stackrel{\text{dist}}{=} T_j(1)$  due to homogeneous parameters), so the repairer always prefers to remain at node  $i$  if  $x_i = 1$ , regardless of the  $x_j$  values at other nodes.

Next, consider case (b), in which  $x_i = 0$  and there is a node  $j \neq i$  with  $x_j = 1$ . In this case, from (12) we have  $\Phi_i^{\text{stay}}(x_i) = 0$ , while from (28) and (29) we can see that  $\Phi_{ij}^{\text{move}}(x_j)$  is strictly positive and greater than  $\Phi_{ij}^{\text{wait}}(x_j)$ . Therefore the repairer chooses to switch to move  $j$  in this case.

Finally, consider case (c), in which  $x_j = 0$  for all  $j \in M$ . In this case the repairer chooses to move to the ‘optimal idling position’,  $i^*$  (or remain if it is already there). Under the stated assumptions, the quantity  $\Psi(i)$  in (15) is the same for all  $i \in M$ , so  $i^*$  can be chosen arbitrarily as any of the nodes in  $M$ . This is a rather trivial case and we will show later in the proof that any action by the repairer is optimal in this situation.

In the next stage of the proof we will show that there exists an optimal policy that makes the same decisions as the index policy. First, we introduce some terminology. Consider two distinct states  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and  $\mathbf{x}' = (i', (x'_1, \dots, x'_m))$ . We say that states  $\mathbf{x}$  and  $\mathbf{x}'$  are *symmetric* if  $x_i = x'_{i'}$  and  $\sum_{j \in M} x_j = \sum_{j \in M} x'_j$ . That is, the total number of failed machines is the same under both states, and the repairer is either at a working machine under both states or at a failed machine under both states. For example, in a problem with 3 machines, we say that states  $(1, (1, 0, 1))$  and  $(2, (0, 1, 1))$  are symmetric, but the states  $(1, (0, 0, 1))$  and  $(2, (0, 1, 0))$  are not symmetric (because the repairer is at a working machine in the first case, but a failed machine in the second case).

Consider a finite-horizon version of the discrete-time MDP formulated in Section 3 (following uniformization), and let  $w_k(\mathbf{x})$  denote the minimum possible expected total cost that can be achieved over  $k$  time steps, given that the initial state is  $\mathbf{x} \in S$ . In order to make progress we will need to use the following three properties of the function  $w_k$ , which we label (WN1)-(WN3) for convenience:

1. (WN1) If states  $\mathbf{x}$  and  $\mathbf{x}'$  are symmetric, then  $w_k(\mathbf{x}) = w_k(\mathbf{x}')$  for all  $k \in \mathbb{N}$ .
2. (WN2) If states  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and  $\mathbf{x}' = (i', (x'_1, \dots, x'_m))$  are identical except that  $x_j = 0$  and  $x'_j = 1$  for some  $j \in M$ , then  $w_k(\mathbf{x}) \leq w_k(\mathbf{x}')$  for all  $k \in \mathbb{N}$ .
3. (WN3) Suppose states  $\mathbf{x} = (i, (x_1, \dots, x_m))$  and  $\mathbf{x}' = (i', (x'_1, \dots, x'_m))$  are identical except that  $i \neq i'$ . If  $x_i = 1$  and  $x'_{i'} = 0$ , then  $w_k(\mathbf{x}) \leq w_k(\mathbf{x}')$  for all  $k \in \mathbb{N}$ .

Property (WN1) states that the optimal expected total cost remains the same if the initial state  $\mathbf{x}$  is replaced by one of its symmetric states. Indeed, if  $\mathbf{x}$  and  $\mathbf{x}'$  are symmetric then  $w_k(\mathbf{x})$  and  $w_k(\mathbf{x}')$  must be identical, since the decision-maker is faced with the same problem in both cases (one may observe that the states  $\mathbf{x}$  and  $\mathbf{x}'$  are identical up to a renumbering of the nodes in  $V$ ). Property (WN2) states that it is better for machine  $j$  to be working than failed (if all else is equal), which is obvious because a failed machine will cause an extra cost to be incurred. Property (WN3) states that it is better for the repairer to be located at a failed machine than a working machine (if all else is equal), which makes sense because being at a failed machine allows the repairer to commence repairs without any delay. All three of these properties can be proved using induction on  $k$ . We omit the induction proofs for brevity.

Next, we consider the action that should be chosen in order to minimize the expected total cost over  $(k + 1)$  time steps, starting from state  $\mathbf{x} = (i, (x_1, \dots, x_m))$ . In general, we have

$$w_{k+1}(\mathbf{x}) = \min_{j \in M} \{q_{k+1}(\mathbf{x}, j)\}, \quad (30)$$

where  $q_{k+1}(\mathbf{x}, j)$  is the expected total cost for choosing action  $j \in M$  and then following an optimal policy for the remaining  $k$  time steps. The expression for  $q_{k+1}(\mathbf{x}, j)$  takes different forms depending on whether the action  $j$  is to remain at the current node  $i$  or to switch to an alternative node. In the former case, it is also useful to distinguish between the cases  $x_i = 0$  and  $x_i = 1$ . Let  $M_0(\mathbf{x})$  and  $M_1(\mathbf{x})$  be the total numbers of machines that are working and failed under state  $\mathbf{x}$ , respectively. Then, using the transition probabilities in (1), we have

$$q_{k+1}(\mathbf{x}, j) = \begin{cases} M_1(\mathbf{x})f(1) + M_0(\mathbf{x})\lambda w_k(\mathbf{x}^{j+}) \\ + (1 - M_0(\mathbf{x})\lambda) w_k(\mathbf{x}), & \text{if } j = i \text{ and } x_i = 0, \\ M_1(\mathbf{x})f(1) + M_0(\mathbf{x})\lambda w_k(\mathbf{x}^{j+}) + \mu w_k(\mathbf{x}^{j-}) \\ + (1 - M_0(\mathbf{x})\lambda - \mu) w_k(\mathbf{x}), & \text{if } j = i \text{ and } x_i = 1, \\ M_1(\mathbf{x})f(1) + M_0(\mathbf{x})\lambda w_k(\mathbf{x}^{j+}) + \tau w_k(\mathbf{x}^{-j}) \\ + (1 - M_0(\mathbf{x})\lambda - \tau) w_k(\mathbf{x}), & \text{if } j \neq i, \end{cases} \quad (31)$$

where  $\mathbf{x}^{j+}$  denotes a state identical to  $\mathbf{x}$  except that the component  $x_j$  is changed from 0 to 1,  $\mathbf{x}^{j-}$  denotes a state identical to  $\mathbf{x}$  except that the component  $x_j$  is changed from 1 to 0, and  $\mathbf{x}^{-j}$  denotes a state identical to  $\mathbf{x}$  except that the repairer's location is changed to  $j$ .

Next, we consider the same three cases (a)-(c) discussed earlier and, for each case, show that the action chosen by the index policy (identified earlier) is also the action that should be chosen in order to minimize the finite-horizon expected total cost. Case (a) is where the repairer is at a machine  $i$  with  $x_i = 1$ . In order to show that remaining at machine  $i$  minimizes the expected total cost over  $k + 1$  stages, we need to show that  $q_{k+1}(\mathbf{x}, i) \leq q_{k+1}(\mathbf{x}, j)$  for all  $j \neq i$ . Using the expressions in (31), it can be checked that

$$q_{k+1}(\mathbf{x}, i) - q_{k+1}(\mathbf{x}, j) = \mu(w_k(\mathbf{x}^{j-}) - w_k(\mathbf{x})) - \tau(w_k(\mathbf{x}^{-j}) - w_k(\mathbf{x})). \quad (32)$$

By Property (WN2) we have  $w_k(\mathbf{x}^{j-}) - w_k(\mathbf{x}) \leq 0$ . If  $j$  is a node with  $x_j = 1$  then  $\mathbf{x}$  and  $\mathbf{x}^{-j}$  are symmetric states, so it follows from (WN1) that  $w_k(\mathbf{x}^{-j}) - w_k(\mathbf{x}) = 0$ . On the other hand, if  $j$  is a node with  $x_j = 0$  then by (WN3) we have  $w_k(\mathbf{x}^{-j}) - w_k(\mathbf{x}) \geq 0$ . It follows that the expression in (32) is non-positive in all cases, so an optimal policy will choose to remain at node  $i$  if  $x_i = 1$ . Next, consider case (b), where the repairer is at a node  $i$  with  $x_i = 0$  and there is at least one node

$j$  with  $x_j = 1$ . In this case, from (31) we obtain

$$q_{k+1}(\mathbf{x}, i) - q_{k+1}(\mathbf{x}, j) = -\tau(w_k(\mathbf{x}^{\rightarrow j}) - w_k(\mathbf{x})), \quad (33)$$

and by (WN3) we have  $w_k(\mathbf{x}^{\rightarrow j}) \leq w_k(\mathbf{x})$ , so the expression in (33) is non-negative. Hence, an optimal policy will choose to switch to node  $j$  in this scenario. Finally, consider case (c), where the repairer is at a node  $i$  with  $x_i = 0$  and we also have  $x_j = 0$  for all  $j \neq i$ . The expression in (33) also applies to this case, but this time (WN1) implies that  $w_k(\mathbf{x}^{\rightarrow j}) - w_k(\mathbf{x})$  is equal to zero because states  $\mathbf{x}$  and  $\mathbf{x}^{\rightarrow j}$  are symmetric. Essentially, this is a trivial case in which any action is optimal, including the action taken by the index policy.

In summary, we have shown that the actions chosen by the index policy in the various possible cases (a)-(c) coincide with the actions chosen by an optimal policy in a finite-horizon problem in which the objective is to minimize total cost over  $k$  stages (for arbitrary  $k \in \mathbb{N}$ ). If we define  $h(\mathbf{x}) := \lim_{k \rightarrow \infty} (w_k(\mathbf{x}) - w_k(\mathbf{z}))$ , where  $\mathbf{z}$  is a fixed reference state in  $S$ , then it follows from standard theory (see Puterman (1994), p. 373) that the function  $h$  satisfies the optimality equations (4). Hence, by our preceding arguments, the actions chosen by the index policy always attain the minimum on the right-hand side in (4), and the index policy is indeed optimal.  $\square$

## D Necessity of the conditions (i)-(iii) in Proposition 4.5

In Proposition 4.5, three conditions are stated for the optimality of the index policy: (i) all machines are directly connected to each other; that is,  $V$  is a complete graph; (ii) each machine has only two possible states; that is,  $K_j = 1$  for each  $j \in M$ ; (iii) the degradation rates, repair rates and cost functions are the same for all machines. In this appendix we consider each of the conditions (i)-(iii) in turn and, for each one, present a counter-example showing that the index policy may not be optimal if the condition is relaxed while the other two conditions are retained.

The results described in examples (a)-(c) below are obtained by using dynamic programming (DP) methods to calculate the optimal average costs  $g^*$  and comparing these with the average costs  $g_{\theta\text{IND}}$  given by the index policy. To obtain the optimal values  $g^*$  we use policy iteration, as specified in Appendix H. To obtain the index policy values  $g_{\theta\text{IND}}$  we simply evaluate the index policy by using the evaluation step in policy iteration, omitting any improvement steps.

Our counter-examples are described below.

- (a) Consider a case where conditions (ii) and (iii) in Proposition 4.5 hold, but condition (i) is violated. We consider a star network as defined in Definition 4.6, with radius  $r = 1$  and  $m = 3$ , and  $K_1 = K_2 = K_3 = 1$ . The degradation rates are  $\lambda_1 = \lambda_2 = \lambda_3 = 0.04$ , the repair rates are  $\mu_1 = \mu_2 = \mu_3 = 0.12$ , and the switching rate is  $\tau = 0.024$ . The cost functions are of type 1 (see Appendix G), with  $c_1 = c_2 = c_3 = 1$ . In this case,  $g_{\theta\text{IND}} = 2.37$ , whereas  $g^* = 2.25$ .
- (b) Consider a case where conditions (i) and (iii) in Proposition 4.5 hold, while condition (ii) is

violated. We consider a system with  $m = 3$  machines, all directly connected to each other, and  $K_1 = K_2 = K_3 = 2$ . The degradation rates are  $\lambda_1 = \lambda_2 = \lambda_3 = 0.089$ , the repair rates are  $\mu_1 = \mu_2 = \mu_3 = 0.52$ , and the switching rate is  $\tau = 0.11$ . The cost functions are of type 1 with  $c_1 = c_2 = c_3 = 1$ . In this case,  $g_{\theta\text{IND}} = 2.62$ , whereas  $g^* = 2.58$ .

(c) Consider a case where conditions (i) and (ii) hold, while condition (iii) is violated. We consider a system with  $m = 3$  machines, all directly connected to each other, and  $K_1 = K_2 = K_3 = 1$ . We describe three subcases:

- (Heterogeneous degradation rates.) We take  $\lambda_1 = 0.034$ ,  $\lambda_2 = 0.16$ , and  $\lambda_3 = 0.055$ . The repair rates are  $\mu_1 = \mu_2 = \mu_3 = 0.74$ , the switching rate is  $\tau = 0.22$ , and the cost functions are of type 1 with  $c_1 = c_2 = c_3 = 1$ . In this case,  $g_{\theta\text{IND}} = 0.85$ , whereas  $g^* = 0.80$ .
- (Heterogeneous repair rates.) We take  $\mu_1 = 0.82$ ,  $\mu_2 = 0.12$ , and  $\mu_3 = 0.63$ . The degradation rates are  $\lambda_1 = \lambda_2 = \lambda_3 = 0.056$ , the switching rate is  $\tau = 0.15$ , and the cost functions are of type 1 with  $c_1 = c_2 = c_3 = 1$ . In this case,  $g_{\theta\text{IND}} = 1.22$ , whereas  $g^* = 1.18$ .
- (Heterogeneous cost parameters.) We consider cost functions of type 1 with  $c_1 = 8.6$ ,  $c_2 = 13.0$ , and  $c_3 = 8.1$ . The degradation rates are  $\lambda_1 = \lambda_2 = \lambda_3 = 0.14$ , the repair rates are  $\mu_1 = \mu_2 = \mu_3 = 0.56$ , and the switching rate is  $\tau = 0.36$ . In this case,  $g_{\theta\text{IND}} = 13.15$ , whereas  $g^* = 12.98$ .

## E Proof of Proposition 4.7

Let  $V$  be a star network as defined in Definition 4.6, with radius  $r \geq 1$ . The other conditions of the proposition are that  $K_j = 1$  for all  $j \in M$ , all  $m$  machines have identical parameter settings, and  $\tau > 2r\lambda$ . The state space has the form

$$S = \{(i, (x_1, \dots, x_m)) \mid i \in V, x_j \in \{0, 1\} \text{ for } j \in M\}.$$

Machine  $i \in M$  is working if  $x_i = 0$ , and has failed if  $x_i = 1$ . As in the proof of Proposition 4.5 we will omit the subscripts on the degradation rates  $\lambda_j$ , repair rates  $\mu_j$  and cost functions  $f_j$ , since these are assumed to be the same for all  $j \in M$ . Also, given that each machine has only two possible conditions, we will depart from our usual notational convention and use  $k$  as a machine or time step index rather than an indicator of a machine's condition.

The proof of Proposition 4.5 used dynamic programming arguments to show that the decisions made by the index policy are optimal, but this approach becomes more complicated in a network with intermediate stages, so instead we will use a slightly different approach. We will show that the decisions made by the index policy are *greedy*, in the sense that they always minimize the expected

amount of time until the repairer is next repairing a machine, and then show that a greedy policy is optimal under the conditions of the proposition.

The first step is to determine the decisions made under the index policy. In the arguments that follow, it will be useful to define  $d(i, j)$  as the length of a shortest path between two nodes  $i, j \in V$ . We can consider five possible cases: (a) the repairer is at a machine  $i \in M$  with  $x_i = 1$ ; (b) the repairer is at a machine  $i \in M$  with  $x_i = 0$  and there is at least one machine  $j \neq i$  with  $x_j = 1$ ; (c) the repairer is at a machine  $i \in M$  with  $x_i = 0$  and we also have  $x_j = 0$  for all  $j \neq i$ ; (d) The repairer is at an intermediate stage  $i \in N$  and there is at least one machine  $j \in M$  with  $x_j = 1$ ; (e) the repairer is at an intermediate stage  $i \in N$  and we have  $x_j = 0$  for all  $j \in M$ .

In case (a), the repairer has a choice between remaining at the failed machine  $i$  or switching to an alternative machine. Recall that the index policy makes decisions using the indices  $\Phi_i^{\text{stay}}(x_i)$ ,  $\Phi_{ij}^{\text{move}}(x_j)$  and  $\Phi_{ij}^{\text{wait}}(x_j)$  defined in (12), (13) and (14) respectively. As in the proof of Proposition 4.5, the index for remaining at node  $i$  simplifies to

$$\Phi_i^{\text{stay}}(x_i) = \Phi_i^{\text{stay}}(1) = \frac{\mathbb{E}[R_i(1)]}{\mathbb{E}[T_i(1)]} = \frac{\mu}{\lambda} f(1). \quad (34)$$

Next, consider the index for moving to an alternative node  $j \neq i$ . According to the rules of the heuristic, the repairer should only consider switching to nodes belonging to the set  $J$  defined in (16). The indices  $\Phi_{ij}^{\text{move}}(x_j)$  and  $\Phi_{ij}^{\text{wait}}(x_j)$  for  $j \neq i$  can be expressed as

$$\Phi_{ij}^{\text{move}}(x_j) = \begin{cases} \mathbb{P}(X_j = 1) \frac{\mathbb{E}[R_j(1)]}{\mathbb{E}[D_{ij}|X_j = 1] + \mathbb{E}[T_j(1)]}, & \text{if } x_j = 0, \\ \frac{\mathbb{E}[R_j(1)]}{\mathbb{E}[D_{ij}] + \mathbb{E}[T_j(1)]}, & \text{if } x_j = 1, \end{cases} \quad (35)$$

$$\Phi_{ij}^{\text{wait}}(x_j) = \frac{\mathbb{E}[R_j(1)]}{1/\lambda + \mathbb{E}[D_{ij}] + \mathbb{E}[T_j(1)]}. \quad (36)$$

It can be seen that if  $x_i = 1$  then  $\Phi_i^{\text{stay}}(x_i) > \Phi_{ij}^{\text{move}}(x_j)$ , regardless of whether or not node  $j$  is included in the set  $J$ . Hence, the repairer will remain at node  $i$  in this case. In case (b), where the repairer is at a node  $i$  with  $x_i = 0$  and we also have  $x_j = 1$  for some  $j \neq i$ , it can be seen from the expressions above that  $\Phi_{ij}^{\text{move}}(x_j) > \Phi_{ij}^{\text{wait}}(x_j)$ , so machine  $j$  is included in the set  $J$  and furthermore the index for remaining at node  $i$  is zero, so the repairer moves towards node  $j$  in this case (this implies taking a step towards the center of the star network).

In case (c), where the repairer is at a machine  $i$  with  $i = 0$  and we also have  $x_j = 0$  for all  $j \neq i$ , it is necessary to identify the node  $i^*$  that minimizes the quantity  $\Psi(i)$  defined in (15). Given the layout of the star network (see Figure 2), it is easy to see that  $\Psi(i)$  is minimized by setting  $i = s$ , where  $s$  is the center node. Hence, in this case (as in the previous case), the repairer should move from machine  $i$  towards the center.

In case (d), where the repairer is at an intermediate stage  $i$  and there exists  $j \in M$  with  $x_j = 1$ ,

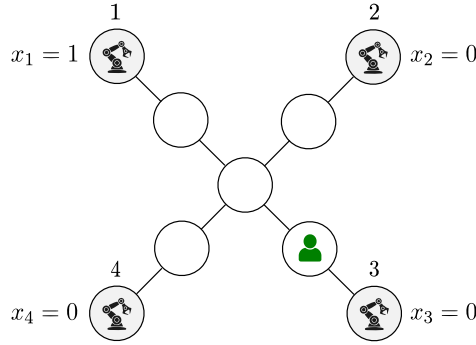


Figure 4: A star network with 4 machines and a radius  $r = 2$ , where the repairer is at an intermediate stage, machine 1 has failed with  $x_1 = 1$  and other machines are working with  $x_2 = x_3 = x_4 = 0$ .

the rules of the heuristic imply that the repairer needs to identify the node  $j$  that minimizes  $\Phi_{ij}^{\text{move}}(x_j)$ . Let  $j$  be a fixed node in  $V$  such that  $x_j = 1$ . If  $1 \leq d(i, j) \leq r$  (implying that the repairer is somewhere between the center node and  $j$ ), then it is clear that  $\mathbb{E}[D_{ij}] \leq \mathbb{E}[D_{ik}]$  for any machine  $k \neq j$  and therefore from (35) we have  $\Phi_{ij}^{\text{move}}(x_j) \geq \Phi_{ik}^{\text{move}}(x_k)$ . Therefore the repairer should move towards node  $j$  in this situation. On the other hand, suppose we have  $1 \leq d(i, k) < r$  for some machine  $k$  such that  $x_k = 0$  (see Figure 4). In this case, the decision is less clear. The repairer could move towards the nearest machine  $k$ , in the hope that a degradation event will occur at that machine before the repairer arrives. However, we will show that the decision under the index policy is to move towards the damaged machine  $j$ , which is further away. In order to show that  $\Phi_{ij}^{\text{move}}(1) \geq \Phi_{ik}^{\text{move}}(0)$ , we begin by establishing some useful bounds. Firstly, we use the general fact that  $1 - x^p \leq k(1 - x)$  for  $k \in (0, 1)$  and  $p \in \mathbb{N}$  to show that

$$\mathbb{P}(X_k = 1) = 1 - \left( \frac{\tau}{\tau + \lambda} \right)^{d(i, k)} \leq \frac{d(i, k)\lambda}{\tau + \lambda}. \quad (37)$$

Notably, this bound becomes tighter as  $\tau/(\tau + \lambda)$  approaches one, which is relevant to our case because we assume  $\tau > 2r\lambda$ . Secondly, we note that  $\mathbb{E}[D_{ik}|X_k = 1]$  is the expected total amount of time for  $d(i, k) + 1$  events to occur, where each event is either a completed switch or a degradation event at node  $k$ , and the latter type of event can only occur once. The expected amount of time for each of these  $d(i, k) + 1$  events is either  $1/(\tau + \lambda)$  or  $1/\tau$ , depending on whether or not the degradation event has occurred yet. Hence,

$$\mathbb{E}[D_{ik}|X_k = 1] \geq \frac{d(i, k) + 1}{\tau + \lambda}. \quad (38)$$

Using the bounds in (37) and (38), we have

$$\begin{aligned}
 \Phi_{ik}^{\text{move}}(0) &= \mathbb{P}(X_k = 1) \frac{\mathbb{E}[R_k(1)]}{\mathbb{E}[D_{ik}|X_k = 1] + \mathbb{E}[T_k(1)]} \\
 &\leq \frac{d(i, k)\lambda}{\tau + \lambda} \cdot \frac{\mathbb{E}[R_k(1)]}{(d(i, k) + 1)/(\tau + \lambda) + \mathbb{E}[T_k(1)]} \\
 &= \frac{d(i, k)\mu f(1)}{d(i, k)\mu + \lambda + \mu + \tau}.
 \end{aligned} \tag{39}$$

On the other hand, using  $E[D_{ij}] = (2r - d(i, k))/\tau$ , the index for switching to machine  $j$  is

$$\Phi_{ij}^{\text{move}}(1) = \frac{\mathbb{E}[R_j(1)]}{\mathbb{E}[D_{ij}] + \mathbb{E}[T_j(1)]} = \frac{\tau\mu f(1)}{\lambda(\tau + (2r - d(i, k))\mu)}. \tag{40}$$

Using (39) and (40), we find (after some algebra steps) that  $\Phi_{ij}^{\text{move}}(1)$  is greater than or equal to the upper bound for  $\Phi_{ik}^{\text{move}}(0)$  provided that

$$\tau^2 + (d(i, k) + 1)\mu\tau \geq (d(i, k) - 1)\lambda\tau + (2rd(i, k) - d(i, k)^2)\lambda\mu. \tag{41}$$

Given that  $\tau > 2r\lambda$ , it is sufficient to show that (41) holds with  $\lambda$  replaced by  $\tau/(2r)$ . After making this substitution and rearranging, we obtain

$$\frac{2r + d(i, k)^2}{2r}\mu\tau + \frac{2r - (d(i, k) - 1)}{2r}\tau^2 \geq 0,$$

which holds because  $1 \leq d(i, k) \leq r$ . In conclusion, under the index policy the repairer chooses to move towards the damaged machine  $j$  rather than the (nearer) undamaged machine  $k$ .

Finally, consider case (e), where the repairer is at an intermediate stage and  $x_j = 0$  for all  $j \in M$ . As noted in case (c), the quantity  $\Psi(i)$  defined in (15) is minimized when  $i = s$ , so the repairer moves towards the center in this situation.

In the next part of the proof, we consider cases (a)-(e) again and show that, in each case, the repairer's action under the index policy also minimizes the expected amount of time until repair. For clarity, we define the 'time until repair' as the amount of time until the repairer chooses to remain at a damaged machine, and abbreviate it as TUR. In case (a), the TUR is trivially equal to zero under the index policy since the repairer remains at the damaged machine  $i$ .

In case (b), the repairer is at a machine  $i$  with  $x_i = 0$  and there exists  $j \neq i$  with  $x_j = 1$ . Suppose (for a contradiction) that the TUR is minimized by remaining at machine  $i$ . If machine  $i$  is still undamaged at the next time step then, logically, the TUR is still minimized by remaining at node  $i$ , since the repairer is still in the same situation. (It may happen that a degradation event occurs at another machine  $k \notin \{i, j\}$ , but this does not affect the minimization of the TUR, since there is no reason for the repairer to prefer machine  $k$  to machine  $j$ ). Continuing this argument, it follows that the TUR is minimized by remaining at node  $i$  until a degradation event occurs there.

However, the expected time until this happens is  $1/\lambda$ , while the expected amount of time to move directly to machine  $j$  is  $2r/\tau$ , and we have  $2r/\tau < 1/\lambda$  by assumption. Therefore, in fact, the repairer should not remain at machine  $i$  and should take the only possible alternative action, which is to move towards the center of the network. Note that we have not shown, at this stage, that the repairer should move to machine  $j$  without any interruption in order to minimize the TUR; we have only shown that if the repairer is at machine  $i$  then it should take the first step towards the center. The actions chosen at intermediate stages will be discussed in case (d).

Next, we address cases (c) and (e) at the same time. In both of these cases we have  $x_j = 0$  for all  $j \in M$ . Let  $i$  denote the repairer's current location, which could be either a machine or an intermediate stage. We note that the quantity  $\Psi(i)$  in (15) is the expected amount of time for the repairer to travel directly towards the next machine that degrades. It can easily be checked that  $\Psi(i)$  increases with  $d(i, s)$ ; that is, the further away the repairer gets from the center, the larger  $\Psi(i)$  becomes. It follows that in order to minimize the TUR the repairer should travel towards the center node  $s$ . The index policy also directs the repairer to move to the center when all nodes are working, so again we find that it minimizes the TUR.

The only remaining case is (d), where the repairer is at an intermediate stage  $i$  and there is at least one node  $j$  with  $x_j = 1$ . If  $d(i, j) \leq r$  then it is obvious that the TUR is minimized by moving to machine  $j$  without any interruption, which the index policy does. Thus, in order to avoid triviality, we should consider again the situation shown in Figure 4, where the repairer's nearest node  $k$  has  $x_k = 0$ . In this case we can use an inductive argument to show that the TUR is minimized by moving towards the center. Initially, suppose  $d(i, k) = 1$ , so the repairer is only one step from machine  $k$ . The TUR cannot be minimized by remaining at the intermediate stage  $i$ , since (using similar reasoning to case (b)) this would imply that the repairer should remain at node  $i$  until a degradation event occurs at machine  $k$ , in which case the TUR is at least  $1/\lambda$ , which is greater than  $(2r - 1)/\tau$  (the TUR given by moving directly to machine  $j$ ). On the other hand, the TUR cannot be minimized by moving to machine  $k$  either, since we already showed in case (b) that if the repairer is at a machine  $k$  with  $x_k = 0$  then it should move towards the center in order to minimize the TUR; therefore, moving from node  $i$  to machine  $k$  would result in back-and-forth movements between nodes  $i$  and  $k$  until a degradation occurs at machine  $k$ , and (again) the TUR would be at least  $1/\lambda$ . Hence, by process of elimination over the action set, we conclude that if  $d(i, k) = 1$  then the TUR is minimized by moving towards the center. An identical argument can be used when  $d(i, k) = 2$ , using the fact that moving towards machine  $k$  results in back-and-forth behavior between two intermediate stages and a TUR of at least  $1/\lambda$ . Continuing this argument, we conclude that the repairer should move towards the center and then continue towards machine  $j$  (or another damaged machine) in order to minimize the TUR.

In summary, we have shown that in all of the cases (a)-(e) the action chosen by the index policy also minimizes the expected time until repair (TUR). The final stage of the proof is to show that a 'greedy' policy that always minimizes the TUR is also optimal. In general, we can think of the evolution of the system as a sequence of time periods, where some of these periods are spent

repairing a machine and some are spent performing other actions (such as moving between nodes in the network). Let these time periods be written as follows:

$$T_1, S_1, T_2, S_2, \dots, T_k, S_k, \dots$$

where each quantity  $S_k$  (for  $k \in \mathbb{N}$ ) is an amount of time spent repairing a machine, and each quantity  $T_k$  is an amount of time spent performing other actions. For example, the repairer might begin at a machine  $i$  with  $x_i = 0$ , but then move towards a damaged machine  $j$  and repair it until it is fixed, in which case  $T_1$  is the amount of time spent moving towards machine  $j$  and  $S_1$  is the time spent performing the repair. The repairer earns a reward of  $(\mu/\lambda)f(1)$  during each of the  $S_k$  periods, and earns no reward during the  $T_k$  periods. Thus, in order to maximize long-run average reward, the ratio  $\sum_k S_k / (\sum_k (S_k + T_k))$  over a long time horizon should be as large as possible. It is clear from this argument that the repairer should never interrupt a repair in progress, since doing so would imply that (possibly) fewer degradation events would occur at other machines during the repair period, and therefore the repairer would not only earn a smaller reward from the current machine (due to interrupting the repair) but would also have fewer opportunities to earn rewards at other machines. We deduce that in order to maximize long-run average reward, the repairer should remain at a damaged machine if it is currently located there.

We will show that in order to maximize long-run average reward, the repairer should always greedily minimize the time until the next repair. This is not necessarily obvious, because if the repairer allows a particular non-repair time  $T_k$  to increase then this might enable the next non-repair time  $T_{k+1}$  to become smaller due to more degradation events occurring during the  $k^{\text{th}}$  non-repair period and more opportunities for earning rewards becoming available.

Let the system be initialized in an arbitrary state and let  $j$  denote the first machine repaired by the index policy, given a particular sample path (i.e. a particular sequence of random events). Hence,  $T_1$  is the time until the repairer begins repairing machine  $j$  (this may be zero if the repairer is initially located at  $j$ ). By our previous arguments,  $\mathbb{E}[T_1]$  is minimized under the index policy, while  $\mathbb{E}[S_1]$  is maximized (because the repair is not interrupted). When the repairer finishes repairing machine  $j$ , there are two possibilities: either (i) all machines are undamaged and the repairer has to wait until the next degradation event occurs, or (ii) there is at least one damaged machine,  $j \neq i$ , in the system. This idea is shown in Figure 5, in which the second non-repair period has been divided into subperiods,  $A_2$  and  $B_2$ . We use  $A_2$  to denote the waiting time until there is at least one damaged machine in the system (which is zero in the case that there is already a damaged machine when the first repair finishes), and  $B_2 = T_2 - A_2$  to denote the additional time until the second repair period begins. In Figure 5 we also use  $t^{\text{deg}}$  to denote the time that the next degradation event occurs, given that there are no damaged machines at time  $T_1 + S_1$ .

Given that we assume repairs are non-interrupted,  $\mathbb{E}[S_1]$  is fixed and minimizing  $\mathbb{E}[T_1]$  corresponds to maximizing  $\mathbb{E}[A_2]$ . If  $A_2 > 0$  then the actions taken by the repairer during the time period  $(T_1 + S_1, T_1 + S_1 + A_2)$  ensure that the minimum TUR starting from time  $T_1 + S_1 + A_2$  (denoted by  $B_2$ ) is itself minimized. Intuitively, because the repairer moves towards the center of

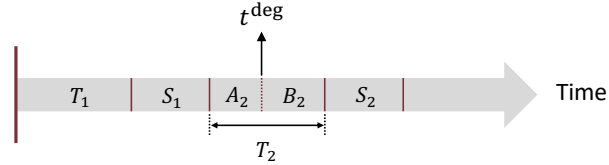


Figure 5: Division of time into periods  $T_1, S_1, T_2, S_2, \dots$  under the index heuristic.

the network while it is waiting for the next degradation event to occur, it is in the best possible position to begin the second repair as early as possible. This ensures that the expected amount of non-repair time before starting the second repair,  $\mathbb{E}[T_1 + T_2]$ , is minimized. We emphasize the subtle point that the index policy does not necessarily minimize  $\mathbb{E}[T_2]$  (because  $T_2$  includes a possible ‘waiting period’  $A_2$ ), but the time spent in the waiting period works to the advantage of the index policy and ensures that the repairer spends as much time as possible moving towards the center, from which it follows that  $\mathbb{E}[B_2]$  and  $\mathbb{E}[T_1 + T_2]$  are both minimized.

This argument can be continued in an inductive manner. In general we can replace  $T_1$  with  $\sum_{k=1}^K T_k$  in Figure 5, where  $K \geq 1$ , and use the inductive assumption that the index policy minimizes  $\mathbb{E}[\sum_{k=1}^K T_k]$ . Then the same arguments as above can be used to show that  $\mathbb{E}[\sum_{k=1}^{K+1} T_k]$  is also minimized. Given that the  $S_k$  values are independent and identically distributed (as noted above), it follows that the index policy maximizes the amount of time spent repairing over any finite time horizon, and therefore maximizes the expected total reward over a finite horizon. Therefore it must also maximize the long-run average reward.  $\square$

## F Details of Online Policy Improvement

In this appendix we provide further details and pseudocode for the online policy improvement (OPI) method described in Section 5. The method consists of an offline part, in which we acquire estimates of the value function  $h_{\theta^{\text{M-IND}}}(\mathbf{x})$  under the modified index policy (referred to as the base policy)  $\theta^{\text{M-IND}}$ , and an online part, in which we apply an improvement step to the base policy (with the base policy also providing an insurance option in cases where an improving action cannot be identified with sufficient confidence). Both parts involve sampling trajectories of possible future events starting from a given state, and we present the pseudocode for this procedure as a separate module, called ‘SampleTrajectory’. We also present the offline part as two separate modules (a preparatory phase and a main phase), and the online part as a single module. In summary, the OPI method is presented here as 4 separate modules:

- SampleTrajectory (called upon by other modules)
- Offline part - preparatory phase
- Offline part - main phase

- Online part

We briefly explain the steps of these modules in subsections F.1-F.4, and provide full pseudocode for all 4 modules in subsection F.5.

To simplify notation, we will denote the modified index policy throughout this appendix by  $\theta \equiv \theta^{\text{M-IND}}$ . The offline and online parts use a dynamic array  $U$  to store and update estimated average cost values  $\hat{h}_\theta(\mathbf{x})$  for states  $\mathbf{x} \in S$ . This array also stores other statistics, including a sum of squared costs  $\widehat{SS}_\theta(\mathbf{x})$ , sum of squared weights  $\widehat{W}_\theta(\mathbf{x})$ , and observation count  $\hat{s}_\theta(\mathbf{x})$  for each  $\mathbf{x} \in U$ . For readability, we drop the hats and  $\theta$ -subscripts in both the discussion and accompanying pseudocode, referring to these statistics as  $h$ -values,  $SS$ -values,  $W$ -values and  $s$ -values.

## F.1 SampleTrajectory Module

The SampleTrajectory module requires certain inputs from the other modules, including a policy for selecting actions  $\theta$ , an estimate for the average cost  $g_\theta$ , an initial state  $\mathbf{z}$ , a reference state  $\mathbf{u}_0$ , an array  $U$  with the latest gathered statistics (as described above) and an integer  $p$  that determines how much information is recorded from each simulated trajectory. It generates a sample trajectory starting from  $\mathbf{z}$  under policy  $\theta$ , and uses this to update the value estimates  $h(\mathbf{x})$  and other statistics for newly encountered states  $\mathbf{x}$ .

Trajectories are sampled by simulating the system evolution over a finite time, recording the cumulative cost  $C$  and the total number of elapsed time steps  $T$ , until a stopping condition is satisfied. This condition is met if the simulated trajectory reaches either (i) a designated reference state  $\mathbf{u}_0$ , or (ii) a state already stored in the value estimate array  $U$ , excluding the initial state  $\mathbf{z}$ . At this point, the simulation terminates and the stopping state is labeled as  $\mathbf{u}$ .

During the simulation, each newly encountered state  $\mathbf{x}$  (i.e. one not previously visited in the same trajectory) is recorded in a `visited` list, along with the cumulative cost  $C_{\mathbf{x}}$  and time step  $T_{\mathbf{x}}$  when  $\mathbf{x}$  was first visited. Here,  $C_{\mathbf{x}}$  represents the total accumulated cost from the start of the trajectory up to the first visit to  $\mathbf{x}$ , and  $T_{\mathbf{x}}$  denotes the number of steps taken to reach  $\mathbf{x}$  from the initial state.

Once the simulated trajectory ends, the algorithm processes the first  $p$  visited states. For each such state  $\mathbf{x}$ , if it is not already included in  $U$ , a new entry  $U(\mathbf{x}) = (h(\mathbf{x}), SS(\mathbf{x}), W(\mathbf{x}), s(\mathbf{x}))$  is created and initially set to  $(0, 0, 0, 0)$ . Otherwise, the current values are extracted directly from  $U$  for update. Each update is computed using the bootstrapped estimate

$$h^{(s)}(\mathbf{x}) \leftarrow (C - C_{\mathbf{x}}) + h(\mathbf{u}) - g_\theta \cdot (T - T_{\mathbf{x}}),$$

where  $s = s(\mathbf{x})$  is the observation count for state  $\mathbf{x}$  and  $h(\mathbf{u})$  denotes the latest available estimate for state  $\mathbf{u}$ , computed using the values  $h^{(r)}(\mathbf{u})$  for  $r = 1, \dots, s(\mathbf{u})$ . This corresponds to equation (19) (please see the accompanying explanation in that section). We then update the estimate  $h(\mathbf{x})$

using the rule

$$h(\mathbf{x}) \leftarrow (1 - \alpha^{(s)}(\mathbf{x}))h(\mathbf{x}) + \alpha^{(s)}(\mathbf{x})h^{(s)}(\mathbf{x}), \quad (42)$$

where  $\alpha^{(s)}(\mathbf{x})$  is a learning parameter, given by

$$\alpha^{(s)}(\mathbf{x}) = \frac{10}{10 + s - 1}.$$

We note that (42) is a standard exponential averaging rule used in RL algorithms, which effectively states that the updated estimate  $h(\mathbf{x})$  is computed as a weighted average of the old estimate and the newly-acquired observation (from the latest simulation). The second moment  $SS(\mathbf{x})$ , and sum of squared weights  $W(\mathbf{x})$  are updated using a rule analogous to (42), and  $U(\mathbf{x})$  is updated accordingly. The module returns the updated array  $U$ , the stopping state  $\mathbf{u}$ , and the total amount of time elapsed during the simulation,  $\tau_\delta$ .

## F.2 Offline part - preparatory phase

In the preparatory phase we estimate the long-run average cost  $g_\theta$  and construct a representative set of system states. This module is divided into three stages:

- (i) We begin by generating a core set of states  $Z^{\text{core}}$ , consisting of one state  $\mathbf{z}_i$  for each machine  $i \in M$ . Specifically, for each  $i \in M$ , we simulate events starting from state  $(i, (0, \dots, 0))$  under the base policy  $\theta$  and identify the state visited most frequently during this simulation, which we denote by  $\mathbf{z}_i$ .
- (ii) Subsequently, a separate long-run simulation from initial state  $(1, (0, \dots, 0))$  is used to estimate the average cost  $g_\theta$ . During this simulation, we track how often the repairer visits the machines in  $M$ . The machine  $j^*$  with the highest visitation count is identified, and its associated core state  $\mathbf{z}_{j^*}$  is designated as the reference state  $\mathbf{u}_0$ . This reference state then becomes the initial member of the array  $U$  mentioned previously.
- (iii) Next, we create a larger set  $Z \supseteq Z^{\text{core}}$  to improve coverage of the state space. For each machine  $i$  we add the core state  $\mathbf{z}_i \in Z^{\text{core}}$  to  $Z$  and also add all states of the form  $\mathbf{z}_i^{\rightarrow j}$  for all nodes  $j$  adjacent to  $i$  in the network, where  $\mathbf{z}_i^{\rightarrow j}$  is identical to  $\mathbf{z}_i$  except that the repairer's location is changed to  $j$ . Also, if  $\mathbf{z}_i$  is a state with  $i^{\text{th}}$  component greater than or equal to one, we include the state  $\mathbf{z}_i^{i-}$  in  $Z$ , where  $\mathbf{z}_i^{i-}$  is identical to  $\mathbf{z}_i$  except that the  $i^{\text{th}}$  component is reduced by one. Including states of the form  $\mathbf{z}_i^{\rightarrow j}$  and  $\mathbf{z}_i^{i-}$  in  $Z$  is beneficial for the online part that follows later.

## F.3 Offline part - main phase

This module takes the set  $Z$  from the preparatory phase as an input. It is divided into two stages:

- (i) We consider each of the states  $\mathbf{z} \in Z$  in turn. For each  $\mathbf{z} \in Z$  we repeatedly sample trajectories of future events starting from  $\mathbf{z}$  by using the SampleTrajectory module, and we update the array  $U$  with the statistics ( $h$ -values,  $SS$ -values,  $W$ -values, and  $s$ -values) gathered from these trajectories. We also set  $p = 1$  within the SampleTrajectory module in this stage. This process continues (returning to the initial state  $\mathbf{z}$  at the end of each trajectory) until either  $R^{\text{off}}$  trajectories have been simulated (where  $R^{\text{off}}$  is a large integer), or the cumulative simulation time exceeds a large threshold  $\tau_{\text{max}}$ . At this point, we consider the next state in  $Z$ , or (if all states in  $Z$  have been considered) move on to the next stage.
- (ii) We consider each of the core states  $\mathbf{z} \in Z^{\text{core}} \subseteq Z$  in turn. For each state  $\mathbf{z} \in Z^{\text{core}}$  we sample a trajectory of future events starting from  $\mathbf{z}$  and update the array  $U$  with the statistics gathered, in the same way as in the previous stage. However, when the trajectory ends (which happens when the system reaches a state  $\mathbf{u} \in U$ ), we use  $\mathbf{u}$  as the initial state for the next trajectory, rather than returning to state  $\mathbf{z}$ . This creates a chain of successive trajectories, with the ending point of one trajectory acting as the starting point for the next. We also set  $p = 5$  within the SampleTrajectory module in this stage. As in the previous stage, this process continues until either  $R^{\text{off}}$  trajectories have been simulated or the cumulative simulation time exceeds the threshold  $\tau_{\text{max}}$ , at which point we consider the next state in  $Z^{\text{core}}$ , or (if all states in  $Z^{\text{core}}$  have been considered) end the offline part.

#### F.4 Online part

The online part is based on the paradigm of a ‘real’ system that we are able to observe and control in a continuous manner over time. The aim is to apply an improvement step to the base policy  $\theta$ , making use of the statistical information gained from simulating the system evolving under  $\theta$  in the offline part. In the absence of being able to observe a ‘real’ system, we conduct a simulation of the system, and perform ‘nested’ simulations within this in order to gather extra information about nearby states, as described in Section 5.

At each iteration, the current system state  $\mathbf{x}$  is observed and a local neighborhood  $F(\mathbf{x})$  is constructed. This neighborhood includes (i) the current state  $\mathbf{x}$  itself, (ii) states reachable by relocating the repairer to adjacent nodes, and (iii) a state resulting from repair completion if the repairer is currently at machine  $i$  with  $x_i \geq 1$ . We then use the empirical data stored in  $U$  to obtain, for each  $\mathbf{y} \in F(\mathbf{x})$ , a confidence interval for  $h_\theta(\mathbf{y})$ . The confidence interval is given by

$$[h^-(\mathbf{y}), h^+(\mathbf{y})] = \left[ h(\mathbf{y}) \pm z_{0.025} \sqrt{\left( \frac{SS(\mathbf{y}) - h(\mathbf{y})^2}{1 - W(\mathbf{y})} \right) W(\mathbf{y})} \right], \quad (43)$$

where  $z_{0.025} \approx 1.96$  is a Normal distribution percentile. The formula in (43) is derived from the theory of weighted sample variances and reliability weights (cf. Sec. 21.7 of Galassi et al. (2011)). In the following lines, we provide some additional details of how the formula is derived. Let  $h(\mathbf{y})$

be the  $h$ -value estimate for state  $\mathbf{y}$  after  $s(\mathbf{y})$  observations have been collected. It can be checked using (42) that

$$h(\mathbf{y}) = \sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y}) h^{(r)}(\mathbf{y}),$$

where the weights  $w^{(r)}(\mathbf{y})$  for  $r = 1, \dots, s(\mathbf{y})$  satisfy

$$w^{(r)}(\mathbf{y}) = \alpha^{(r)}(\mathbf{y}) \prod_{j=r+1}^{s(\mathbf{y})} (1 - \alpha^{(j)}(\mathbf{y})).$$

We can compute the sample variance as

$$\begin{aligned} \hat{\sigma}(\mathbf{y})^2 &= \frac{\sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y}) (h^{(r)}(\mathbf{y}) - h(\mathbf{y}))^2}{\sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y})} \\ &= \sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y}) \cdot h^{(r)}(\mathbf{y})^2 - h(\mathbf{y})^2 \\ &= SS(\mathbf{y}) - h(\mathbf{y})^2. \end{aligned}$$

Then, to obtain an unbiased estimate  $\sigma(\mathbf{y})$ , we use

$$\sigma(\mathbf{y})^2 = \frac{\hat{\sigma}(\mathbf{y})^2}{1 - (V_2/V_1^2)},$$

where

$$V_1 = \sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y}), \quad V_2 = \sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y})^2.$$

As noted above,  $V_1 = 1$  and we also have  $V_2 = W(\mathbf{y})$ . Hence,

$$\sigma(\mathbf{y})^2 = \frac{SS(\mathbf{y}) - h(\mathbf{y})^2}{1 - W(\mathbf{y})}.$$

Thus, the effective variance of the weighted sample mean is

$$\frac{\sigma(\mathbf{y})^2}{1/\sum_{r=1}^{s(\mathbf{y})} w^{(r)}(\mathbf{y})^2} = \left( \frac{SS(\mathbf{y}) - h(\mathbf{y})^2}{1 - W(\mathbf{y})} \right) W(\mathbf{y}),$$

which justifies (43). After computing the confidence intervals, the algorithm selects an action  $a^* \in A_{\mathbf{x}}$  by, firstly, checking to see whether there is an action  $a^*$  satisfying the inequality

$$\sum_{\mathbf{y} \in S} p_{\mathbf{x},\mathbf{y}}(a^*) h(\mathbf{y}) < \sum_{\mathbf{y} \in S} p_{\mathbf{x},\mathbf{y}}(a') h(\mathbf{y})$$

for all alternative actions  $a' \neq a^*$  and for all possible values of  $h(\mathbf{y})$  (for  $\mathbf{y} \in F(\mathbf{x})$ ) within their

respective confidence intervals  $[h^-(\mathbf{y}), h^+(\mathbf{y})]$ . If no such  $a^*$  exists, we resort to the base policy as a ‘safety measure’ and set  $a^* = \theta(\mathbf{x})$ .

To continuously improve the accuracy of the value estimates, the SampleTrajectory module is called at every iteration. On each iteration, it is allowed to run for a maximum  $\delta = 0.01$  seconds and the statistics in  $U$  are updated during this period. The motivation for this step is that in the ‘real’ system, there would be a certain amount of time ( $\delta$ ) in between any pair of state transitions, and therefore we can make use of this time to improve value function estimates for states that are near the current state. Finally, each iteration ends with a real-time transition from the current state  $\mathbf{x}$  to a successor state based on the chosen action  $a^*$  (note that this is the action given by the ‘improving’ step, not the action given by the base policy). After a large number of iterations,  $R^{\text{on}}$ , have been completed, the online part ends and the average cost over all time steps is returned as a performance metric for the OPI policy.

## F.5 Pseudocode of OPI

On the next four pages we provide pseudocode for the following modules, which are used as part of the online policy improvement method:

- SampleTrajectory
- Offline part - preparatory phase
- Offline part - main phase
- Online part

Please note that full descriptions of these modules (including their purposes, inputs and outputs) are provided in Appendices F.1, F.2, F.3 and F.4 respectively.

**SampleTrajectory module**


---

```

1: Input: State  $\mathbf{z}$ ; reference state  $\mathbf{u}_0$ ; array  $U$ ; recording length  $p$ ; policy  $\theta$ ; average cost  $g_\theta$ 
2: Initialize:
3: Total cost incurred  $C \leftarrow 0$ ; total time steps  $T \leftarrow 0$ 
4: Current state  $\mathbf{z}_{\text{curr}} \leftarrow \mathbf{z}$ ; visited list  $\text{visited} \leftarrow [(\mathbf{z}, 0, 0)]$ 
5: Start time:  $t_{\text{start}} \leftarrow$  current time
6: while true do
7:    $C \leftarrow C + c(\mathbf{z}_{\text{curr}})$ ;  $T \leftarrow T + 1$ 
8:   Simulate transition:  $\mathbf{z}_{\text{curr}} \xrightarrow{\theta(\mathbf{z}_{\text{curr}})} \mathbf{x}$ 
9:   if  $(\mathbf{x} \neq \mathbf{z}$  or  $\mathbf{x} = \mathbf{u}_0)$  and  $\mathbf{x} \in U$  then
10:     Set stopping state  $\mathbf{u} \leftarrow \mathbf{x}$ ; break
11:   else
12:      $\mathbf{z}_{\text{curr}} \leftarrow \mathbf{x}$ 
13:     if  $\mathbf{x} \notin \text{visited}$  then
14:       Append  $(\mathbf{x}, C, T)$  to  $\text{visited}$ 
15:     end if
16:   end if
17: end while
18: for each  $(\mathbf{x}, C_x, T_x)$  in  $\text{visited} [ : p]$  do ▷ first  $p$  visited during simulation
19:   if  $\mathbf{x} \notin U$  then
20:     Initialize:  $U(\mathbf{x}) \leftarrow (0, 0, 0, 0)$ 
21:   end if
22:   Unpack:  $(h(\mathbf{x}), SS(\mathbf{x}), W(\mathbf{x}), s(\mathbf{x})) \leftarrow U(\mathbf{x})$ ;  $(h(\mathbf{u}), SS(\mathbf{u}), W(\mathbf{u}), s(\mathbf{u})) \leftarrow U(\mathbf{u})$ 
23:    $s \leftarrow s(\mathbf{x}) + 1$ ;  $s(\mathbf{x}) \leftarrow s$ 
24:   Compute learning parameter:  $\alpha^{(s)}(\mathbf{x}) \leftarrow \frac{10}{10+s-1}$ 
25:   Compute estimate:  $h^{(s)}(\mathbf{x}) \leftarrow (C - C_x) + h(\mathbf{u}) - g_\theta \cdot (T - T_x)$ 
26:   Update:
      
$$h(\mathbf{x}) \leftarrow (1 - \alpha^{(s)}(\mathbf{x})) \cdot h(\mathbf{x}) + \alpha^{(s)}(\mathbf{x}) \cdot h^{(s)}(\mathbf{x})$$

      
$$SS(\mathbf{x}) \leftarrow (1 - \alpha^{(s)}(\mathbf{x})) \cdot SS(\mathbf{x}) + \alpha^{(s)}(\mathbf{x}) \cdot (h^{(s)}(\mathbf{x}))^2$$

      
$$W(\mathbf{x}) \leftarrow (1 - \alpha^{(s)}(\mathbf{x}))^2 \cdot W(\mathbf{x}) + (\alpha^{(s)}(\mathbf{x}))^2$$

27:    $U(\mathbf{x}) \leftarrow (h(\mathbf{x}), SS(\mathbf{x}), W(\mathbf{x}), s(\mathbf{x}))$ 
28: end for
29: End time:  $t_{\text{end}} \leftarrow$  current time; simulation time  $\tau_\delta \leftarrow t_{\text{end}} - t_{\text{start}}$ 
30: Output: Array  $U$ ; stopping state  $\mathbf{u}$ ; simulation time  $\tau_\delta$ 
Note: For each state  $\mathbf{x}$ ,  $s(\mathbf{x})$  is the number of estimates so far. The current estimate  $h^{(s)}(\mathbf{x})$  and learning parameter  $\alpha^{(s)}(\mathbf{x})$  correspond to the  $s$ -th update.

```

---

**Offline part - preparatory phase**

- 
- 1: **Input:** Base policy  $\theta$ ; simulation lengths  $R^{(1)}, R^{(2)}$
  - 2: **Stage 1: Construct core state set  $Z^{\text{core}}$**
  - 3: Initialize  $Z^{\text{core}} \leftarrow \emptyset$
  - 4: **for** each machine  $i \in M$  **do**
  - 5:     Initialize  $r \leftarrow 0$ ,  $\mathbf{x} \leftarrow (i, (0, \dots, 0))$ ,  $S_i \leftarrow \emptyset$ ; set  $N_i(\mathbf{x}) \leftarrow 0$  upon adding  $\mathbf{x}$  to  $S_i$
  - 6:     **while**  $r < R^{(1)}$  **do**
  - 7:         Simulate transition:  $\mathbf{x} \xrightarrow{\theta(\mathbf{x})} \mathbf{y}$
  - 8:          $\mathbf{x} = (j, (x_1, \dots, x_m)) \leftarrow \mathbf{y}$ ;  $r \leftarrow r + 1$
  - 9:         **if**  $j = i$  **then**
  - 10:              $S_i \leftarrow S_i \cup \{\mathbf{x}\}$ ;  $N_i(\mathbf{x}) \leftarrow N_i(\mathbf{x}) + 1$
  - 11:         **end if**
  - 12:     **end while**
  - 13:      $\mathbf{z}_i \leftarrow \arg \max_{\mathbf{x} \in S_i} N_i(\mathbf{x})$ ;  $Z^{\text{core}} \leftarrow Z^{\text{core}} \cup \{\mathbf{z}_i\}$
  - 14: **end for**
  - 15: **Stage 2: Estimate average cost  $g_\theta$  and choose reference state  $\mathbf{u}_0$**
  - 16: Initialize  $C \leftarrow 0$ ,  $r \leftarrow 0$ ,  $\mathbf{x} \leftarrow (1, (0, \dots, 0))$ , and counters  $n_i \leftarrow 0$  for all  $i \in M$
  - 17: **while**  $r < R^{(2)}$  **do**
  - 18:      $C \leftarrow C + c(\mathbf{x})$
  - 19:     Simulate transition:  $\mathbf{x} \xrightarrow{\theta(\mathbf{x})} \mathbf{y}$
  - 20:      $\mathbf{x} = (i, (x_1, \dots, x_m)) \leftarrow \mathbf{y}$ ;  $r \leftarrow r + 1$
  - 21:     **if**  $i \in M$  **then**
  - 22:          $n_i \leftarrow n_i + 1$
  - 23:     **end if**
  - 24: **end while**
  - 25:  $g_\theta \leftarrow C/R^{(2)}$ ;  $j^* \leftarrow \arg \max_i n_i$ ;  $\mathbf{u}_0 \leftarrow \mathbf{z}_{j^*}$ ; move  $\mathbf{z}_{j^*}$  to the front of  $Z^{\text{core}}$
  - 26: **Stage 3: Construct representative state set  $Z$**
  - 27: Initialize  $Z \leftarrow \emptyset$
  - 28: **for** each  $\mathbf{z}_i = (i, (z_1, \dots, z_m)) \in Z^{\text{core}}$  **do**
  - 29:      $Z \leftarrow Z \cup \{\mathbf{z}_i\} \cup \{\mathbf{z}_i^{\rightarrow j} : j \text{ adjacent to } i\}$
  - 30:     **if**  $z_i \geq 1$  **then**
  - 31:          $Z \leftarrow Z \cup \{\mathbf{z}_i^{i-}\}$
  - 32:     **end if**
  - 33: **end for**
  - 34: **Output:** Average cost  $g_\theta$ ; reference state  $\mathbf{u}_0$ ; core state set  $Z^{\text{core}}$ ; representative state set  $Z$
-

---

**Offline part - main phase**

---

- 1: **Input:** Average cost  $g_\theta$ ; reference state  $\mathbf{u}_0$ ; core state set  $Z^{\text{core}}$ ; representative state set  $Z$ ; simulation length  $R^{\text{off}}$ ; time limit  $\tau_{\text{max}}$
  - 2: Initialize array  $U \leftarrow \emptyset$ ; set  $U(\mathbf{u}_0) \leftarrow (0, 0, 1, 1)$
  - 3: **Stage 1: Value estimation for representative states**
  - 4: **for** each  $\mathbf{z} \in Z$  **do**
  - 5:     Initialize  $r \leftarrow 0$ ;  $\tau \leftarrow 0$ ; recording length  $p \leftarrow 1$
  - 6:     **while**  $r < R^{\text{off}}$  **and**  $\tau < \tau_{\text{max}}$  **do**
  - 7:         Run **SampleTrajectory**( $\mathbf{z}$ )  $\rightarrow U$ , stopping state  $\mathbf{u}$ , time  $\tau_\delta$  ▷ Ignore  $\mathbf{u}$
  - 8:          $r \leftarrow r + 1$ ;  $\tau \leftarrow \tau + \tau_\delta$
  - 9:     **end while**
  - 10: **end for**
  - 11: **Stage 2: System evolution from core states**
  - 12: **for** each  $\mathbf{z} \in Z^{\text{core}}$  **do**
  - 13:     Initialize  $r \leftarrow 0$ ;  $\tau \leftarrow 0$ ; recording length  $p \leftarrow 5$
  - 14:     **while**  $r < R^{\text{off}}$  **and**  $\tau < \tau_{\text{max}}$  **do**
  - 15:         Run **SampleTrajectory**( $\mathbf{z}$ )  $\rightarrow U$ , stopping state  $\mathbf{u}$ , time  $\tau_\delta$
  - 16:          $\mathbf{z} \leftarrow \mathbf{u}$ ;  $r \leftarrow r + 1$ ;  $\tau \leftarrow \tau + \tau_\delta$
  - 17:     **end while**
  - 18: **end for**
  - 19: **Output:** Array  $U$
-

**Online part**

- 
- 1: **Input:** Average cost  $g_\theta$ ; reference state  $\mathbf{u}_0$ ; array  $U$ ; simulation length  $R^{\text{on}}$ ;  $\delta$ -length time limit
  - 2: Initialize  $C \leftarrow 0$ ,  $r \leftarrow 0$ ,  $\mathbf{x} \leftarrow \mathbf{u}_0$
  - 3: **while**  $r < R^{\text{on}}$  **do**
  - 4:      $C \leftarrow C + c(\mathbf{x})$
  - 5:     **Stage 1: Action selection**
  - 6:     Define local neighborhood  $F(\mathbf{x})$
  - 7:     **for each**  $\mathbf{y} \in F(\mathbf{x})$  **do**
  - 8:         Compute confidence interval  $[h^-(\mathbf{y}), h^+(\mathbf{y})]$  from  $U$
  - 9:     **end for**
  - 10:     Select  $a^* \in A_{\mathbf{x}}$  satisfying:
 
$$\sum_{\mathbf{y} \in S} p_{\mathbf{x}, \mathbf{y}}(a^*) h(\mathbf{y}) < \sum_{\mathbf{y} \in S} p_{\mathbf{x}, \mathbf{y}}(a') h(\mathbf{y}), \quad \forall a' \neq a^*, \quad \forall h(\mathbf{y}) \in [h^-(\mathbf{y}), h^+(\mathbf{y})].$$
  - 11:     **if** no such  $a^*$  exists **then**
  - 12:         Set  $a^* \leftarrow \theta(\mathbf{x})$
  - 13:     **end if**
  - 14:     **Stage 2: Information supplement**
  - 15:     Initialize  $\tau \leftarrow 0$ ; recording length  $p \leftarrow 1$
  - 16:     **while**  $\tau < \delta$  **do**
  - 17:         Simulate transition:  $\mathbf{x} \xrightarrow{a^*} \mathbf{x}'$
  - 18:         Define local neighborhood  $F(\mathbf{x}')$
  - 19:         **for each**  $\mathbf{y} \in F(\mathbf{x}')$  **do**
  - 20:             Run **SampleTrajectory**( $\mathbf{y}$ )  $\rightarrow U, \mathbf{u}, \tau_\delta$
  - 21:              $\tau \leftarrow \tau + \tau_\delta$
  - 22:         **end for**
  - 23:     **end while**
  - 24:     **Stage 3: Real-time transition**
  - 25:     Simulate transition:  $\mathbf{x} \xrightarrow{a^*} \mathbf{x}_{\text{new}}$
  - 26:      $\mathbf{x} \leftarrow \mathbf{x}_{\text{new}}$ ;  $r \leftarrow r + 1$
  - 27: **end while**
  - 28:  $g \leftarrow C/R^{\text{on}}$
  - 29: **Output:** Average cost  $g$

**Note:** Confidence intervals are computed using the method of reliability weights (Galassi et al. (2011), Sec. 21.7)

---

## G Methods for generating the parameters for the numerical experiments in Section 6

For each of the 1,000 instances considered in Section 6, the system parameters are randomly generated as follows:

- The type of cost function is sampled unbiasedly from the set  $\{1, 2, 3\}$ , where:
  - (i) Type 1 represents a linear cost function, i.e., the cost incurred per unit time for machine  $i \in M$  in state  $x_i$  is given by  $f_i(x_i) = c_i x_i$ ;
  - (ii) Type 2 represents a quadratic cost function, i.e., the cost incurred per unit time for machine  $i \in M$  in state  $x_i$  is given by  $f_i(x_i) = c_i x_i^2$ ;
  - (iii) Type 3 represents a piecewise linear cost function, i.e., the cost incurred per unit time for machine  $i \in M$  in state  $x_i$  is given by  $f_i(x_i) = c_i (x_i + 10 \cdot \mathbb{I}(x_i = K_i))$ , where  $\mathbb{I}(x_i = K_i)$  is the indicator function that equals 1 if  $x_i = K_i$ , and 0 otherwise.
- The ‘failed’ state,  $K$ , is sampled unbiasedly from the set  $\{1, 2, 3, 4, 5\}$ .
- The number of machines,  $m$ , is sampled unbiasedly from the set  $\{2, 3, 4, 5, 6, 7, 8\}$ .
- The  $a$ -coordinate and  $b$ -coordinate of each machine  $i \in M$ , denoted as  $(a_i, b_i)$ , are independently sampled unbiasedly from the set  $\{1, 2, 3, 4, 5\}$ , with resampling used if two or more machines share the same pair of coordinates.
- The machines in  $M$  are re-numbered according to the  $a$ -coordinate first and then (in case of ties) by the  $b$  coordinate.
- The overall traffic intensity,  $\rho$ , is sampled from a continuous uniform distribution between 0.1 and 1.5. Subsequently, the degradation rates  $\lambda_i$  and repair rates  $\mu_i$  for  $i \in M$  are generated as follows:
  - Each repair rate  $\mu_i$  is initially sampled from a continuous uniform distribution between 0.1 and 0.9.
  - For each machine  $i \in M$ , an initial value for the degradation rate  $\lambda'_i$  is sampled from a continuous uniform distribution between  $0.1\mu_i$  and  $\mu_i$ .
  - For each machine  $i \in M$ , the actual traffic intensity  $\rho_i$  is obtained by re-scaling the initial traffic intensity  $\lambda'_i/\mu_i$ , as follows:

$$\rho_i := \frac{\lambda'_i/\mu_i}{\sum_{i \in M} \lambda'_i/\mu_i} \rho.$$

This ensures that  $\sum_{i \in M} \rho_i = \rho$ .

- For each machine  $i \in M$ , the actual degradation rate  $\lambda_i$  is obtained as follows:

$$\lambda_i := \rho_i \mu_i.$$

- All of the  $\lambda_i$  and  $\mu_i$  values are rounded to 2 significant figures (also causing a small change to the value of  $\rho$ ).
- For each machine  $i \in M$ , the cost rate,  $c_i$ , is sampled from a continuous uniform distribution between 0.1 and 0.9.
- In order to generate the switching rate  $\tau$ , we first define  $\eta := \tau / (\sum_{i \in M} \lambda_i)$  and generate the value of  $\eta$  as follows:
  - Sample a value  $p$  from a continuous uniform distribution between 0 and 1.
  - If  $p < 0.5$ , sample  $\eta$  from a continuous uniform distribution between 0.1 and 1.
  - If  $p \geq 0.5$ , sample  $\eta$  from a continuous uniform distribution between 1 and 10.

We then define  $\tau := \eta \sum_{i \in M} \lambda_i$ .

## H Policy Iteration Algorithm for systems with $2 \leq m \leq 4$

The following pseudocode shows how policy iteration is used to calculate the optimal value of  $g^*$  for systems with  $2 \leq m \leq 4$  in our numerical experiments. Specifically, we implement the policy iteration algorithm described in Puterman (1994) (p. 378), with the relative value function for each

candidate policy computed via successive approximations until a convergence criterion is met.

---

### Policy Iteration Algorithm

---

- 1: **Input:** base policy  $\theta$ , reference state  $\mathbf{x}_0$ , tolerance  $\varepsilon$
  - 2: **Initialization:**
  - 3:  $v(\mathbf{x}) \leftarrow 0, g(\mathbf{x}) \leftarrow 0$  for all  $\mathbf{x} \in \theta$
  - 4:  $v^+(\mathbf{x}) \leftarrow 0, g^+(\mathbf{x}) \leftarrow 0$  for all  $\mathbf{x} \in \theta$
  - 5:  $g^{\text{error}} \leftarrow \infty$
  - 6: **Stage 1: Policy Evaluation Algorithm (PEA)**
  - 7: **while**  $g^{\text{error}} \geq \varepsilon$  **do**
  - 8:      $g^{\text{error}} \leftarrow 0$
  - 9:     **for** each state  $\mathbf{x} \in \theta$  **do**
  - 10:         Compute  $p_{\mathbf{x},\mathbf{y}}$  and cost  $c(\mathbf{x})$
  - 11:         
$$v^+(\mathbf{x}) \leftarrow c(\mathbf{x}) + \sum_{\mathbf{y}} p_{\mathbf{x},\mathbf{y}} v(\mathbf{y}) + \left(1 - \sum_{\mathbf{y}} p_{\mathbf{x},\mathbf{y}}\right) v(\mathbf{x})$$
  - 12:          $g^+(\mathbf{x}) \leftarrow v^+(\mathbf{x}) - v(\mathbf{x})$
  - 13:          $g^{\text{error}} \leftarrow \max\{g^{\text{error}}, |g^+(\mathbf{x}) - g(\mathbf{x})|\}$
  - 14:     **end for**
  - 15:      $v \leftarrow v^+, g \leftarrow g^+$
  - 16: **end while**
  - 17:  $g^\theta \leftarrow g(\mathbf{x}_0)$
  - 18:  $v(\mathbf{x}) \leftarrow v(\mathbf{x}) - v(\mathbf{x}_0)$
  - 19: **Stage 2: Successive Policy Improvement Algorithm (nPPIA)**
  - 20:  $n \leftarrow 0, \theta^n \leftarrow \theta$
  - 21: **repeat**
  - 22:     **for** each state  $\mathbf{x} \in \theta^n$  **do**
  - 23:         For each admissible action  $a$  at  $\mathbf{x}$ , compute
 
$$Q(\mathbf{x}, a) = c(\mathbf{x}) + \sum_{\mathbf{y}} p_{\mathbf{x},\mathbf{y}}(a) v(\mathbf{y}) + \left(1 - \sum_{\mathbf{y}} p_{\mathbf{x},\mathbf{y}}(a)\right) v(\mathbf{x})$$
  - 24:          $\theta^{n+1}(\mathbf{x}) \leftarrow \arg \min_a Q(\mathbf{x}, a)$
  - 25:     **end for**
  - 26:     Evaluate  $\theta^{n+1}$  using **Stage 1 (PEA)**
  - 27:      $n \leftarrow n + 1$
  - 28: **until**  $\theta^n = \theta^{n-1}$
  - 29:  $g^* \leftarrow g(\mathbf{x}_0)$
  - 30: **Output:**  $g^*$
-

Computational times for this algorithm (categorized according to the value of  $m$ ) are reported in Table 12 below.

Table 12: Average running times (in seconds) for policy iteration algorithm, over 412 problem instances with  $2 \leq m \leq 4$  on randomly-generated networks.

	$m = 2$ [132 instances]	$m = 3$ [148 instances]	$m = 4$ [132 instances]
Av. run times	0.091	1.21	11.73

## I Simulation methods for the numerical experiments in Section 6

In each of the 1,000 problem instances, the performances of the index policy and the policies given by the polling heuristic are estimated by simulating the discrete-time evolution of the uniformized MDP described in Section 3. We also use a ‘common random numbers’ method to ensure that machine degradations occur at the same times under each of these policies. More specifically, the simulation steps are as follows:

1. Generate a set of random system parameters as described in Appendix G.
2. Set  $R^{\text{sim}} := 500,000$  as the number of time steps in the simulation.
3. Generate a list  $Z$  of length  $R^{\text{sim}}$ , consisting of uniformly-distributed random numbers between 0 and 1.
4. Consider each heuristic policy in turn. For each one, we set  $\mathbf{x}_0 := (1, (0, 0, \dots, 0))$  as the initial state, and then use the random numbers in  $Z$  to simulate events during these  $R^{\text{sim}}$  time steps, with the system beginning in state  $\mathbf{x}_0$ . The statistics collected during these time steps are used to quantify the heuristic’s performance.

When implementing the OPI policy, we need to decide on the values of  $R^{(1)}$ ,  $R^{(2)}$ ,  $R^{\text{off}}$ ,  $\tau_{\max}$  and  $R^{\text{on}}$  used in the offline and online stages (see the pseudocode in Appendix F for details). We use the values  $R^{(1)} = 10,000$ ,  $R^{(2)} = 500,000$ ,  $R^{\text{off}} = 100,000$ ,  $\tau_{\max} = 100$  and  $R^{\text{on}} = 500,000$ .